UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ÉTUDE LONGITUDINALE DU NIVEAU DE SÉVÉRITÉ D'EXAMINATEURS EN FRANÇAIS LANGUE ÉTRANGÈRE

THÈSE
PRÉSENTÉE
COMME EXIGENCE PARTIELLE
DU DOCTORAT EN ÉDUCATION

PAR
CHRISTOPHE CHÉNIER

NOVEMBRE 2018

REMERCIEMENTS

Merci à monsieur Gilles Raîche, qui a un jour, au restaurant, prononcé les mots fatidiques : « il est possible d'estimer le niveau de sévérité des évaluateurs... ». Tel le visage d'Hélène, cela allait lancer mille galères sur les eaux troubles de l'exploration dont je sors tout juste, encore ébahi par tout le chemin parcouru. Son savoir méthodologique et son enthousiasme de chercheur sont des modèles dignes d'émulation. Merci à monsieur Pascal Ndinga, pour la qualité de son accompagnement et sa jovialité. Je tiens aussi à remercier les autres membres du jury, soit madame Marthe Hurteau, ayant assuré la présidence, ainsi que messieurs Léon Harvey et Éric Dionne. Je remercie également le personnel du centre de test d'où proviennent les données de cette thèse.

À mes parents, qui ont eu recours à la méthode par osmose afin de nourrir ma curiosité intellectuelle : mettre les livres en présence des enfants, en espérant que le transfert s'opère de l'un à l'autre, d'une façon ou d'une autre. Je remercie de même tous mes collègues, chercheurs curieux que j'ai eu le plaisir de côtoyer durant ces années, de même que les étudiants que j'ai eu le plaisir d'avoir dans les cours que j'ai donnés.

À mes amis, à mon frère pour son excellent soutien, à ma cousine et, pour terminer, aux musiciens et compositeurs dont les créations ont nourri ces longues heures de rédaction.

DÉDICACE

À Réjean Ducharme, *il miglior fabbro*

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

La majorité des évaluations à forts enjeux dans le domaine de l'évaluation des langues ont des épreuves d'expression orale ou écrite devant être évaluée par des examinateurs. Or, le recours à des examinateurs introduit un risque potentiel quant à la validité de l'évaluation, car les examinateurs peuvent avoir des niveaux de sévérité différents les uns des autres, ce qui peut affecter négativement les notes accordées aux candidats. De plus, les recherches ayant étudié l'efficacité des formations pour réduire les différences de niveau de sévérité d'examinateurs ont donné des résultats mitigés. L'idée de former et certifier des examinateurs suppose que leur niveau de sévérité soit suffisamment stable pour qu'il ne change pas trop d'une séance d'évaluation à une autre, sans quoi il est impossible d'assurer la qualité des examinateurs à long ou moyen terme. Des études montrent que le niveau de sévérité d'un examinateur peut changer significativement d'un temps de mesure à un autre, mais ces études ont toutes peu de temps de mesure (≤ 12). Afin de combler l'absence d'étude longitudinale du niveau de sévérité, cette thèse cherche donc à « Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du temps chronologique et du nombre de candidats évalués », « Comparer l'évolution du niveau de sévérité des examinateurs débutants et expérimentés » et « Comparer l'évolution du niveau de sévérité d'examinateurs travaillant ensemble ». Les données proviennent d'un centre de test pour l'examen d'expression orale du TEF, examen passé par 3 333 candidats d'octobre 2010 à avril 2014. Les performances ont été évaluées avec une grille analytique de 12 critères d'évaluation. Un total de 12 examinateurs ayant évalué au moins 100 candidats durant cette période ont été retenus pour les analyses. Leur niveau de sévérité a été estimé à l'aide du modèle de Rasch à multifacettes et les séries chronologiques de ces niveaux de sévérité ont été modélisées avec la modélisation AMMI (Autorégressif à Moyenne Mobile Intégré).

Les résultats montrent que 73 % des séries chronologiques des niveaux de sévérité sont normalement distribuées et que 38 % des séries sont autocorrélées. Un peu plus d'un tiers des examinateurs ont un niveau de sévérité dont l'étendue intraindividuelle longitudinale est égale ou supérieure à l'étendue interindividuelle des niveaux de sévérité de l'ensemble des examinateurs au cours d'une période de temps donnée : leur niveau de sévérité a un problème de dérive temporelle. Quelques résultats montrent que certains examinateurs débutants ont un niveau de sévérité plus instable, susceptible d'avoir des valeurs extrêmes par rapport au niveau de leurs collègues expérimentés. De même, dans certains cas, les niveaux de sévérité d'examinateurs travaillant ensemble sont corrélés, ces corrélations allant de faibles à modérées. Ces résultats mènent à une discussion sur les difficultés méthodologiques et conceptuelles de l'étude de l'évolution longitudinale du niveau de sévérité, principalement sur la modélisation temporelle à adopter et sur la nature conceptuelle de ce qu'est la sévérité des examinateurs. Finalement, soulignons que ces résultats proviennent de données secondaires et que, étant donné l'échantillonnage non aléatoire et les particularités de

l'examen du TEF, ils ne sont pas généralisables ; ces limites mènent à la proposition de pistes de recherches futures.

Mots clefs : sévérité des examinateurs, dérive temporelle de la sévérité, français langue étrangère (FLE), séries chronologiques, AMMI

ABSTRACT

Most high-stakes assessments in the field of foreign language testing have writing or speaking performances marked by human raters, but the very presence of human raters poses a potential threat to the assessment's validity. It is well-known that raters can have very different levels of severity, which can then have adverse effects on the ratings testing candidates receive. Furthermore, research shows that rater training aimed at reducing differences between raters' severity levels has mixed effects. The very idea of training raters and certifying their competency supposes that their severity level is sufficiently stable so as to not change too much from one rating session to another; otherwise, it would be difficult to make sure raters are performing adequately over time. Studies have shown that raters' severity level can change significantly from one time to another, but these studies all have a limited number of measurement times (≤ 12). In order to fill the lack of longitudinal studies about the evolution of raters' severity levels, this dissertation will "model the evolution of raters' severity levels according to chronological time and the number of candidates rated", "compare the evolution of novice and experienced raters" and "compare the evolution of the severity level of raters working together". The data come from the speaking tests that took place at a single TEF (*Test d'Évaluation de Français*) test center between October 2010 and April 2014. A total of 3,333 candidates took the test. The candidates were assessed by 12 raters, who each assessed at least 100 candidates, and the candidates were rated with an analytic rubric containing 12 criteria. The raters' severity levels were estimated by using the Many-Facet Rasch Measurement Model and those severity levels were then modeled with ARIMA procedures for time series (Auto-Regressive Integrated Mobile Average).

Results show that 73 % of severity level time series are normally distributed and that 38 % are autocorrelated. Over a third of raters have a severity level whose intraindividual range is equal to or greater than the interindividual range of severity levels of all the examiners for a given time period: in other words, they suffer from severity drift. Results show that certain novice raters have a more unstable severity level, one that is liable to contain extreme values when compared to the severity levels of their more experienced colleagues. There are also cases of raters working together whose severity level time series are correlated, with correlations ranging from weak to moderate. A discussion ensues about the methodological and conceptual difficulties inherent to longitudinal studies of the evolution of severity levels, the problems being mainly about the choice of time modelling and the nature of "severity". Due to non-random sampling and the specific characteristics of the TEF speaking test, this dissertation's results cannot be generalized. These limitations lead to some suggestions of possible future studies.

Keywords: rater severity, severity drift, French as a foreign language (FFL), time series, ARIMA

INTRODUCTION

Un test ou un examen doit, comme tout instrument de mesure, être utilisé de manière appropriée afin que le but poursuivi puisse être atteint. Qu'il s'agisse d'un examen certificatif, d'une évaluation formative, d'un questionnaire de recherche ou d'un inventaire visant à poser un diagnostic psychosocial, tout instrument de ce type devrait mener à une décision et à des actions valides. Depuis la publication des premiers *Standards* en 1954, l'importance du concept de « validité » pour tout test, examen ou questionnaire est reconnue par la communauté de chercheurs et de praticiens en éducation et en psychologie et dans tous les domaines où ce type d'instruments est utilisé (American Psychological Association, American Educational Research Association et National Council on Measurement Used in Education, 1954). Concept protéiforme apparu sous ce nom en 1921, où il désigne « la détermination de ce qu'un test mesure[1] », la validité est essentielle à tout instrument de mesure (Buckingham, McCall, Otis, Rugg, Trabue et Courtis, 1921). Longtemps considérée comme propriété de l'instrument lui-même, la validité est plutôt conçue, depuis les années 80-90, comme propriété de l'ensemble du processus de mesure, de la conceptualisation des besoins des usagers jusqu'aux conséquences de l'utilisation des mesures obtenues (Bachman, 2005 ; Kane, 2013 ; Messick, 1995 ; Newton et Shaw, 2014 ; Xi, 2010).

Selon cette conception dominante de la validité, chaque composante du processus de mesure se doit d'avoir une validité minimale pour que les mesures puissent être utilisées à bon escient. Les composantes incluent entre autres la définition des besoins

[1] Traduction libre.

des utilisateurs de l'instrument et du construit à mesurer, la rédaction des items ou des tâches d'évaluation, le processus de correction, la méthode de notation et les décisions prises à partir des mesures. Les recommandations exactes changent selon les auteurs, mais, pour s'assurer de la validité du processus, il faut justifier la valeur et la qualité de chacune des composantes à l'aide d'arguments et de preuves factuelles.

Parmi toutes ces composantes, le recours à des examinateurs[2] pour évaluer les performances et productions complexes fait l'objet de recherches depuis la fin du 19e siècle (Edgeworth, 1888, 1890). Très tôt, le problème de la subjectivité des examinateurs a été identifié comme une source de problèmes potentiels à la qualité d'un examen, ce que l'on nommerait aujourd'hui sa validité. Ainsi, l'inconstance des examinateurs utilisés lors des examens oraux est l'une des raisons expliquant l'abandon d'examens oraux universitaires, héritiers des *disputationes* médiévales, en faveur d'examens écrits au courant du 19e siècle (Stray, 2001). De même, le rôle et le fonctionnement des examinateurs et les problèmes connexes sont parmi les sujets étudiés avec soin par les fondateurs de la docimologie française, Piéron et Laugier (Barbier, 1983 ; Laugier et Weinberg, 1927 ; Martin, 2002). De très nombreuses recherches ont ainsi été menées sur la subjectivité des examinateurs, son impact et les façons d'y remédier. S'inscrivant dans cette lignée, cette thèse s'intéressera aux examinateurs, plus particulièrement aux différences dans les notes accordées à des performances d'un niveau similaire et à l'évolution temporelle de ces différences.

Le premier chapitre développe la problématique. Le rôle central des examinateurs pour bien des examens à forts enjeux est présenté, de même que les problèmes posés

---

[2] Traduction libre de *rater*. Voir le glossaire.

par les différences de niveau de sévérité entre les examinateurs. La question générale de recherche met fin au chapitre.

Le deuxième chapitre présente la recension des écrits sur la question de la sévérité des examinateurs dans le domaine de l'évaluation en langue. Le concept de sévérité et le modèle de mesure retenu pour l'opérationnaliser sont définis, puis les résultats des études antérieures montrent les écarts de sévérité existants et les quelques recherches ayant étudié la dérive temporelle de la sévérité[3] sont détaillées. Suivent une synthèse des études recensées et, pour clore le chapitre, les objectifs spécifiques de recherche.

Le troisième chapitre concerne la méthodologie. Les sources des données, l'instrument et le test utilisé pour collecter les données sont détaillés. Le traitement des données et les étapes analytiques sont présentés en détail et les analyses précises permettant de répondre à chacune des questions spécifiques sont énumérées, de même que les logiciels utilisés. Les considérations éthiques mettent fin au chapitre.

Le quatrième chapitre contient les résultats, qui sont énumérés en 6 ensembles distincts, en fonction du découpage temporel retenu. Pour chaque découpage temporel, les résultats sont présentés selon les objectifs spécifiques de recherche. La discussion suit au cinquième chapitre et, finalement, la conclusion vient au sixième chapitre. Cette conclusion débute avec un bref rappel du contenu de la thèse, se poursuit en énumérant les limites de celle-ci et se termine avec des pistes de recherches futures.

---

[3] Traduction libre de *severity drift*. Voir le glossaire.

CHAPITRE I

PROBLÉMATIQUE

Bien que peu présents dans le système scolaire québécois, les examens à forts enjeux[4] tendent, depuis 25 ans, à occuper une place de plus en plus importante dans la société québécoise. Un examen à forts enjeux est une épreuve qu'un candidat doit réussir et qui, en cas d'échec, entraîne des conséquences importantes directes, immédiates et négatives pour celui-ci, que ces conséquences soient financières, scolaires, académiques ou encore socioprofessionnelles (Cole et Osterlind, 2008 ; Ryan, 2002 ; Stobart et Eggen, 2012). L'épreuve ministérielle uniforme en langue d'enseignement au collégial (ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie, 2013) est, à notre connaissance, le seul exemple d'examen à forts enjeux dans le système scolaire québécois allant du primaire au collégial, car c'est la seule épreuve dont la réussite est obligatoire pour l'obtention d'un diplôme. D'autres examens à forts enjeux existent toutefois, que ce soit les examens standardisés de compétences langagières, obligatoires au Québec depuis décembre 2011 pour les candidats à l'immigration (ministère de l'Immigration, Diversité et Inclusion, 2018), les examens d'accès à un ordre professionnel (Libman, 2009) ou les évaluations de la performance pour l'embauche, la promotion ou la rémunération (Kline et Sulsky, 2009 ; McKinley et Boulet, 2004).

Plusieurs examens à forts enjeux ont au moins une tâche complexe devant être évaluée par un examinateur, par exemple une expression écrite, une simulation

[4] Voir le glossaire.

clinique ou une étude de cas. Dans ces cas, la performance des candidats est notée par un ou plusieurs examinateurs, ce qui est un souci constant pour les responsables des examens, car les différences potentielles entre la sévérité du jugement de différents examinateurs peuvent affecter les notes attribuées aux candidats et ainsi fausser leurs résultats. Les plus anciennes recherches publiées en éducation relevaient déjà le problème de l'impact potentiel de l'identité des examinateurs sur la note accordée à une performance (Edgeworth, 1888, 1890) et ce problème n'a cessé, depuis, de préoccuper les chercheurs du domaine et les responsables des examens faisant appel au jugement d'examinateurs (Bejar, 2012 ; Kline et Sulsky, 2009 ; Landy et Farr, 1980 ; Myford, 2012 ; Myford et Wolfe, 2003, 2004 ; Saal, Downey et Lahey, 1980). C'est que tout examen, quel qu'en soit le domaine, doit pouvoir mener à une note et à une décision qui sont valides. Or, les modèles théoriques développés depuis les 40 dernières années sur la validité des évaluations en éducation et en psychologie s'accordent sur le fait que, pour être considérées valides, la note et la décision prise à partir de cette note ne doivent dépendre que de ce qui est évalué et de la difficulté des tâches d'évaluation (American Educational Research Association, American Psychological Association et National Council on Measurement in Education, 2014). Toutes les autres caractéristiques pouvant influencer la note et la décision représentent des sources de variance illégitimes et devraient théoriquement être neutralisées. Ainsi, le moment de l'évaluation, le genre ou l'ethnie des candidats et des examinateurs, leur expérience professionnelle, les consignes et instructions données aux candidats, les conditions matérielles et environnementales de passation de l'examen et les interactions entre toutes ces caractéristiques ne devraient avoir aucune influence sur la note et la décision qui en découle.

Malheureusement, plusieurs recherches ont montré que ce n'est pas toujours le cas et que certaines de ces caractéristiques ont parfois une influence indue sur les résultats des candidats, influence assez forte pour que la décision prise suite à l'examen ne soit

pas la bonne (Casanova et Demeuse, 2011 ; Eckes, 2005 ; Engelhard Jr. et Myford, 2003 ; Koretsovska, 2013 ; Lim, 2009 ; McManus, Thompson et Mollon, 2006 ; Yen, Ochieng, Michaels et Friedman, 2005). S'il est relativement facile de neutraliser l'impact des caractéristiques matérielles et environnementales en standardisant les conditions de passation des examens, la neutralisation des caractéristiques des examinateurs est beaucoup plus difficile. Il est bien sûr impossible de standardiser des êtres humains et leur jugement de la même manière que l'on peut standardiser des procédures ou des éléments matériels. Les différences entre les jugements des examinateurs font en sorte qu'une même performance peut être notée différemment, sans que les raisons motivant ces différences soient connues. L'idéal selon lequel les examinateurs devraient être objectifs et interchangeables n'est en pratique pas réalisable et la subjectivité du jugement évaluatif ne peut être éliminée. La recherche a ainsi identifié une vingtaine de biais introduits dans le jugement évaluatif par la subjectivité des examinateurs (Myford et Wolfe, 2003, 2004 ; Saal *et al.*, 1980). Conscients que ces biais sont rarement volontaires, certains chercheurs ont, dès les années 60, utilisé une appellation plus neutre, « effets de l'examinateur[5] », et celle-ci est maintenant l'expression la plus répandue (Desai, 1965 ; Han, 2016 ; Myford et Wolfe, 2003 ; Norman et Goldberg, 1966 ; Wesolowski, 2016). Ce texte utilisera exclusivement cette appellation.

Ces effets de l'examinateur peuvent prendre plusieurs formes : éviter de donner des notes trop fortes ou trop faibles (effet de tendance centrale), noter un critère en fonction d'un autre critère (effet de halo), noter généreusement une performance moyenne parce qu'elle suit une performance très faible (effet de séquence), etc. Les études ont montré que la prévalence des effets de l'examinateur est élevée et que le jugement des examinateurs est régulièrement affecté par l'un ou l'autre de ces effets (Cai, 2012 ; DeCarlo, Kim et Johnson, 2011 ; Hoskens et Wilson, 2001 ; Y.-H. Kim,

[5] Voir le glossaire.

2009 ; Wolfe et McVay, 2010 ; Wolfe, Myford, Engelhard Jr. et Manalo, 2007). De tous les effets de l'examinateur recensés par la recherche, les différences de sévérité ou de clémence représentent l'effet le plus souvent identifié par la recherche. Ces différences de sévérité ou de clémence, qui seront définies plus avant dans le contexte théorique, peuvent être conçues comme le fait de donner, en moyenne, des notes plus élevées ou plus basses que d'autres examinateurs pour une même performance (Myford et Wolfe, 2003). Les expressions « différences de sévérité », « sévérité d'examinateur » ou « niveau de sévérité » seront utilisées indistinctement dans le reste du texte. Les différences de sévérité ont été observées chez des examinateurs dans tous les domaines, incluant l'éducation, l'évaluation en langues étrangères, l'évaluation des performances professionnelles et l'évaluation pour l'accès à un ordre professionnel. Ces différences, lorsqu'elles sont importantes et que les examinateurs sont très sévères ou très cléments, ont un impact sur les notes accordées aux candidats car, contrairement aux autres effets de l'examinateur, les différences de sévérité affectent en moyenne toutes les performances évaluées. Il s'agit de l'effet de l'examinateur ayant un impact sur le plus grand nombre de notes accordées par les examinateurs, ce qui en fait l'effet le plus menaçant pour la validité des examens à forts enjeux (Cronbach, 1990 ; Wang et Yao, 2013 ; Wilson et Case, 1997 ; Wolfe et McVay, 2010, 2012 ; Wolfe *et al.*, 2007). Les notes accordées lors d'examens à forts enjeux et les décisions prises avec ces notes sont à risque d'être invalidées par cet effet de l'examinateur et tout organisme responsable de ce type d'examen doit tenir compte de ce risque toujours présent afin de s'en prémunir.

Traditionnellement, les organisations responsables des examens à forts enjeux ont tenté de neutraliser les différences de sévérité en sélectionnant les examinateurs en fonction de leurs diplômes et de leurs expériences professionnelles, en les formant à leur rôle d'examinateur et en surveillant leurs différences de sévérité pendant les évaluations. Malheureusement, ces garde-fous ne réussissent pas toujours à contrer

les différences de sévérité et plusieurs recherches font état de différences de sévérité ayant un impact important sur les notes attribuées aux candidats et sur les décisions prises à partir de ces notes – par exemple, un échec attribué à un candidat dû à la sévérité de l'examinateur (Eckes, 2005 ; Fuller, Homer, Pell et Hallam, 2016 ; McManus *et al.*, 2006 ; Wolfe *et al.*, 2007). La recherche montre que l'éducation et l'expérience professionnelle ne sont pas corrélées aux différences de sévérité, pas plus que les caractéristiques sociodémographiques des examinateurs (Davis, 2012 ; Eckes, 2005 ; Hsieh, 2011 ; Kachchaf et Solano-Flores, 2012 ; H. J. Kim, 2011 ; Y.-H. Kim, 2009 ; Korenovska, 2013 ; Wei et Llosa, 2015). Les facteurs explicatifs du niveau de sévérité postulés par divers modèles théoriques (Eckes, 2011 ; H. J. Kim, 2011 ; Wolfe et Song, 2015) n'ont pas été vérifiés par la recherche. La capacité de la formation au rôle d'examinateur et de la surveillance en temps réel à réduire suffisamment les différences de sévérité est également ambiguë, de nombreuses études ayant montré une relative impuissance de la formation à éliminer les différences de sévérité ou à les réduire suffisamment pour que ces différences aient un impact minime, acceptable sur les notes des candidats (Davis, 2016 ; Elder, Barkhuizen, Knoch et von Randow, 2007 ; H. J. Kim, 2011 ; Kondo, 2010 ; Weigle, 1994, 1998). De plus, l'efficacité de la formation n'est pas systématique et les explications du succès ou de l'insuccès de la formation ne sont pas connues. Ce flou est particulièrement préoccupant pour les responsables de la formation des examinateurs, car il est difficile d'améliorer cette formation lorsque les raisons de la réussite ou de l'échec de ce processus sont mal connues. Les différences de sévérité, en plus d'être une menace constante à la validité des examens à forts enjeux, restent encore aujourd'hui un sujet de recherche actif (Han, 2016 ; Mallinson, Pape et Guernon, 2016 ; Wu et Tan, 2016).

Il y a toutefois un problème important avec l'état des connaissances sur les différences de sévérité. Ce qui est connu à ce sujet provient presque exclusivement de

recherches transversales ayant mesuré la sévérité d'examinateurs à un seul temps de mesure ou, tout au plus, à 2 ou 3 temps séparés de quelques semaines, voire de quelques jours ou heures. Or, les moyens mis en place par les organisations pour contrer les différences de sévérité n'ont de sens que dans la mesure où le niveau de sévérité d'un examinateur est relativement temporellement stable. C'est seulement à ce titre qu'une intervention visant à modifier un niveau de sévérité jugé inacceptable peut avoir une certaine efficacité non éphémère. Il ne sert à rien de vouloir modifier le niveau de sévérité des examinateurs si ce niveau est essentiellement instable, erratique et fluctue naturellement au fil du temps, car une amélioration du niveau de sévérité suite à une intervention pourrait bien être immédiatement suivie de niveaux ayant des valeurs inacceptables. Le fait de sélectionner les examinateurs en fonction d'un processus de certification ne vaut que si le niveau de sévérité des examinateurs est relativement stable. Sinon, un examinateur pourrait faire preuve d'un niveau de sévérité adéquat au moment de l'examen de certification et ensuite voir son niveau dériver vers des extrêmes inacceptables. Parallèlement, les études sur les facteurs explicatifs des différences de sévérité ne valent que si ces différences peuvent être expliquées, du moins en théorie.

Le fait est que toutes ces actions et ces études partagent le postulat implicite en vertu duquel les différences de sévérité sont autre chose que du bruit aléatoire et qu'il y a bien un objet pouvant être étudié ou modifié. La véracité de ce postulat est toutefois inconnue. Seulement deux études longitudinales ayant mesuré les différences de sévérité à différents moments sur une période continue de plusieurs mois ont été faites jusqu'à maintenant, et, dans les deux cas, les résultats montrent que les différences de sévérité de certains examinateurs fluctuent largement d'un mois à l'autre (Lim, 2009 ; Park, 2011). Ces deux études ont toutefois utilisé un nombre restreint d'examinateurs (de 6 à 18) ou de temps de mesure (de 5 à 12), ce qui en limite la portée. Il est impossible d'affirmer, à partir des résultats de ces deux seules

études, si les différences de sévérité constituent bel et bien un objet stable et, de là, modifiable.

Cette recherche vise par conséquent à répondre à la question de recherche suivante : « Comment évolue longitudinalement le niveau de sévérité d'examinateurs? » Cette recherche contribuera indirectement à l'amélioration de la formation des examinateurs pour les examens à forts enjeux en aidant à établir si, oui on non, les différences de sévérité sont une caractéristique relativement stable des examinateurs et, par conséquent, susceptible d'être corrigée par une formation ciblée. Cette contribution sera toutefois à long terme, puisque cette étude est exploratoire et descriptive. Sur le plan scientifique, cette recherche aidera à savoir s'il est possible d'identifier des facteurs explicatifs des différences de sévérité et de l'évolution temporelle du niveau de sévérité.

CHAPITRE II

CONTEXTE THÉORIQUE

Afin de pouvoir étudier systématiquement le niveau de sévérité d'examinateurs et répondre partiellement à la question de recherche ayant clos la problématique, il faut en premier lieu une définition conceptuelle de ce qu'est le niveau de sévérité d'examinateurs et une manière d'opérationnaliser ce concept. Ensuite, à la lumière du concept opérationnalisé, les résultats pertinents des études antérieures permettront de voir en quoi des réponses partielles ont été apportées à la question de recherche générale. La grande majorité des études antérieures ayant été faites en évaluation en langue, cette thèse s'inscrira dans ce champ et le contexte théorique présentera exclusivement les résultats des études faites dans ce domaine.

La première partie du contexte théorique présentera différentes conceptions du niveau de sévérité. Une définition conceptuelle sera choisie et le modèle de mesure communément utilisé dans la littérature pour estimer le niveau de sévérité sera présenté et expliqué en détail. Ensuite, les résultats concernant le niveau de sévérité d'examinateurs en langue seront présentés. Premièrement, la prévalence et l'ampleur des divergences de niveau de sévérité seront décrites. Deuxièmement, les facteurs langagiers, formationnels et expérientiels potentiels des différences de niveau de sévérité seront résumés. Troisièmement, les résultats directement en lien avec la dérive temporelle de la sévérité et les caractéristiques méthodologiques de ces études seront détaillés, à la fois pour les études ayant utilisé des examinateurs en expression écrite et orale. Finalement, une synthèse générale des résultats ayant trait au niveau de

sévérité et à son évolution temporelle sera faite et les manques de connaissances ainsi identifiés mèneront à la formulation des objectifs spécifiques de recherche de cette thèse.

2.1 Sévérité des examinateurs : concept

Si l'expression « niveau de sévérité » va de soi et est intuitivement comprise par tous, une définition conceptuelle rigoureuse est plus difficile à établir. D'ailleurs, la plupart des études portant sur la sévérité des examinateurs n'en présentent aucune définition conceptuelle étoffée (par exemple : Wang et Yao, 2013). Pour certains auteurs, le niveau de sévérité renvoie à une caractéristique d'un examinateur qui est sévère ou clément de manière stable et constante. Cet examinateur accorde des notes supérieures ou inférieures aux notes que devrait recevoir une performance, de manière systématique et indépendante du domaine d'évaluation (DeCoths, 1977 ; Guilford, 1954 ; Schriesheim, Kinicki et Schriesheim, 1979). Cette conception a toutefois le problème majeur de supposer que la vraie note que mérite une performance peut être connue, ce qui n'est pas le cas, puisque l'inférence quant à la valeur de la vraie note est toujours faite à partir d'un nombre fini de performances, ce qui introduit une erreur de mesure due à l'échantillonnage des performances observées. D'autres auteurs conçoivent la sévérité comme le fait d'attribuer des notes qui sont, en moyenne, inférieures à la note moyenne de l'échelle d'appréciation[6] utilisée (Bernardin, LaShells, Smith et Alvares, 1976 ; Taylor et Hastman, 1956). Cette conception a le problème évident de postuler que la moyenne des résultats des performances évaluées se situe exactement à la moyenne de l'échelle d'appréciation, ce qui est une supposition injustifiable. Des échantillons provenant de populations dont le niveau moyen de performance diffère peuvent néanmoins être évalués à l'aide de la même échelle d'appréciation sans que cela ne pose problème. Or, il se pourrait ainsi que l'un des échantillons ait une moyenne différant de la moyenne de l'échelle

[6] Voir le glossaire.

d'appréciation sans que cet écart à la moyenne ne soit préjudiciable à l'utilisation de cette échelle pour évaluer cet échantillon.

Ces deux conceptions sont rarement retenues aujourd'hui et la plupart des auteurs contemporains conçoivent plutôt le niveau de sévérité comme le fait d'accorder des notes inférieures ou supérieures aux autres examinateurs, lorsque le niveau d'habileté des candidats et le niveau de difficulté des tâches sont pris en compte et contrôlés (Myford et Wolfe, 2003). Il s'agit d'une conception statistique et relative qui ne s'applique qu'à un ensemble délimité d'évaluations pour lesquelles des données suffisantes sont disponibles. Un examinateur est sévère ou clément par rapport aux facettes d'une situation d'évaluation, soit les candidats, les critères d'évaluation, les tâches d'évaluation ou tout autre élément jugé pertinent. Contrairement à la première conception présentée, où un examinateur est supposé sévère quel que soit le domaine d'évaluation (langue, patinage artistique), cette conception suppose qu'un examinateur est sévère de manière circonstancielle. Il est possible que le niveau de sévérité d'un examinateur change en fonction de la situation d'évaluation. Un enseignant pourrait ainsi être sévère en français et clément en univers social. Cette conception suppose que le niveau de sévérité est non pas un construit au sens classique du terme (Cronbach et Meehl, 1955), mais bien une « variable intermédiaire » au sens de MacCorquodale et Meehl (1948) et qu'elle n'est que le résumé de relations empiriques entre divers concepts, soit le niveau d'habileté et le niveau de difficulté de la tâche d'évaluation. Cette conception est agnostique quant à la possibilité que le niveau de sévérité soit ou non un trait stable d'un examinateur, les deux possibilités étant compatibles avec une telle conception. Un examinateur ayant un niveau de sévérité élevé dans une situation d'évaluation donnée pourrait ainsi avoir un niveau de sévérité aussi élevé dans une autre situation, avec de nouveaux candidats et de nouvelles tâches d'évaluation où, au contraire, son niveau

de sévérité pourrait différer. C'est la conception que retiendra initialement cette thèse, dont les résultats permettront de revenir sur la conception de la sévérité.

2.2 Modèle de Rasch à multifacettes

La plupart des études faites sur la sévérité des examinateurs souscrivent à la conception présentée ci-dessus, ce qui suppose qu'elles utilisent un modèle de mesure permettant de mesurer les différents éléments pertinents d'une situation d'évaluation, soit, au minimum, le niveau d'habileté des candidats et le niveau de sévérité des examinateurs. Le modèle de Rasch à multifacettes a été développé dans ce but par Linacre (1994) et il a été largement adopté par la communauté de l'évaluation en langue (par exemple : Eckes, 2009). Ce modèle est une extension du modèle de Rasch de base pour données dichotomiques (Rasch, 1960). Le modèle de Rasch de base permet d'estimer la probabilité qu'un candidat réussisse un item en fonction de la différence entre son niveau d'habileté et le niveau de difficulté de l'item. Ces deux paramètres sont ainsi mesurés sur une même échelle de mesure, soit l'échelle en *logit*. Le modèle de Rasch pour données dichotomiques est représenté par l'équation suivante :

$$P(x_{ni} = 1 \mid \theta, \beta) = \frac{e^{\theta_n - \beta_i}}{1 + e^{\theta_n - \beta_i}} \qquad (1)$$

où θ_n représente l'habileté du candidat n, β_i la difficulté de l'item i, e la base du logarithme naturel et P la probabilité de bonne réponse ($x = 1$) du candidat n à l'item i. Il est possible de constater que, indépendamment de la valeur de l'échelle utilisée, si les paramètres θ et β ont la même valeur, le résultat de l'équation est 0,5, ce qui est la probabilité qu'a un candidat d'avoir la bonne réponse à un item dont le niveau de difficulté est égal à son niveau d'habileté. Le terme « logit » vient de ce que l'équation 1 peut être réécrite sous la forme logarithmique de la manière suivante :

$$\ln\left[\frac{P_{nix=1}}{P_{nix=0}}\right] = \theta_n - \beta_i \quad (2)$$

où ln représente le logarithme naturel du rapport des probabilités de bonne et de mauvaise réponse. Cette reformulation met en évidence la nature additive de la relation entre les paramètres d'habileté du candidat et de difficulté de l'item, ceux-ci étant mesurés sur une échelle commune. Chaque paire de valeurs possibles pour ces paramètres correspond à une probabilité de bonne réponse. Tel que vu précédemment, si la différence de valeur entre ces deux paramètres est de 0, la probabilité de bonne réponse est de 0,5. Si la différence est de 1, la probabilité est de 0,5 ± 0,23 selon que le niveau d'habileté est supérieur de 1 (P = 0,73) ou inférieur de 1 (P = 0,27) au niveau de difficulté. Si la différence est de 1,4, la probabilité est de 0,5 ± 0,30, etc. Le modèle multifacettes, lui, est une extension du modèle de base en ce qu'il est possible d'ajouter tout paramètre jugé pertinent à la situation d'évaluation (chaque paramètre est une facette), ce qui se traduit, pour une situation d'évaluation où le jugement rendu par l'examinateur est « succès » ou « échec », par l'équation suivante :

$$\ln\left[\frac{P_{nijx=1}}{P_{nijx=0}}\right] = \theta_n - \beta_i - \alpha_j \quad (3)$$

où β représente la difficulté de la tâche d'évaluation i et α le niveau de sévérité de l'examinateur j. Typiquement, pour l'évaluation en langue et en éducation, les examinateurs n'accordent pas des notes dichotomiques (« succès » ou « échec »), mais plutôt des notes polychotomiques ordonnées en niveaux de performance (p. ex. les niveaux représentés par les lettres E à A). Puisqu'une échelle polychotomique n'est rien d'autre qu'une série d'échelles dichotomiques (il y a $c - 1$ dichotomies, où c égale le nombre de catégories de l'échelle polychotomique), il est possible d'utiliser le même modèle de base, en ajoutant toutefois un paramètre supplémentaire indiquant la difficulté relative de chaque catégorie, soit le paramètre τ. Le modèle de Rasch à multifacettes pour données polychotomiques peut être représenté par l'équation suivante :

$$\ln\left[\frac{P_{nijx=c}}{P_{nijx=c-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_c \qquad (4)$$

où ln représente le logarithme du rapport des probabilités qu'a un candidat n de se voir assigner la note c plutôt que $c - 1$ par l'examinateur j à la tâche i et où τ représente la difficulté relative de chaque seuil entre deux catégories consécutives.

Chaque élément de chaque facette est mesuré sur la même échelle en logit et ils sont tous directement comparables, mais la valeur de cette échelle est arbitraire. Le point 0 n'a pas de signification précise et il est défini par l'analyste, généralement de manière à correspondre à la valeur moyenne des éléments de cette facette. L'orientation des échelles est également choisie par l'analyste. Pour la sévérité, si l'échelle est orientée négativement (indiquée par une soustraction dans l'équation 4), une valeur positive correspondra à un examinateur plus sévère. Mais si l'échelle était orientée positivement, une valeur positive correspondrait à un examinateur plus clément[7]. De plus, les résultats d'analyses différentes ne sont pas directement comparables. Par exemple, il n'est pas possible d'affirmer qu'un examinateur d'une première étude qui a un niveau de sévérité de 1,7 logit est plus sévère qu'une examinatrice d'une deuxième étude dont le niveau de sévérité est de -2, logits, car le point 0 de chaque étude est arbitraire et diffère d'une étude à l'autre, exactement comme les points 0 des échelles Celsius et Fahrenheit ne dénotent pas la même température. Il faut que certains éléments (candidats ou examinateurs) soient communs à ces deux études pour pouvoir comparer deux examinateurs sur la même échelle. En revanche, que deux études partagent ou non une même échelle, l'interprétation du niveau de sévérité est la même en termes de probabilité qu'une note soit assignée dans une catégorie ou une autre. Précisons que dans une situation d'analyse réelle, les valeurs des paramètres sont inconnues et doivent être estimées à l'aide d'un estimateur cherchant

[7] Sauf mention contraire, tous les résultats présentés dans cette thèse sont orientés négativement. Plus une valeur est élevée, plus l'examinateur est sévère.

les valeurs maximisant la vraisemblance des données réellement obtenues et que le modèle présenté est le modèle *rating scale*, qui suppose que toutes les tâches sont évaluées à l'aide de la même échelle ayant les mêmes seuils.

2.3 Sévérité des examinateurs : recherches antérieures

De nombreuses études ont rapporté des niveaux de sévérité préoccupants, que ce soit pour l'évaluation de l'expression écrite ou orale, comme le montrent les 39 recherches identifiées qui ont mesuré et rapporté les niveaux de sévérité d'examinateurs à l'aide du modèle de Rasch à multifacettes (Bachman, Lynch et Mason, 1995 ; Bonk et Ockey, 2003 ; Brown, 1995 ; Caban, 2003 ; Cai, 2012 ; Casanova et Demeuse, 2011 ; Congdon et McQueen, 2000 ; Davis, 2016 ; Du et Brown, 2000 ; Eckes, 2005 ; Eckes, 2012 ; Elder *et al.*, 2007 ; Engelhard Jr., 1994 ; Engelhard et Myford, 2003 ; Eszter, 2007 ; Fahim et Bijani, 2011 ; Hsieh, 2011 ; Johnson et Lim, 2009 ; Kassim, 2007 ; H. J. Kim, 2011 ; Knoch, Fairbairn et Huisman, 2015 ; Kondo, 2010 ; Kondo-Brown, 2002 ; Korenovska, 2013 ; Lim, 2009 ; Lopes Toffoli, de Andrade et Cezar Bornia, 2016 ; Lumley et McNamara, 1995 ; Meier, 2014 ; O'Loughlin, 2002 ; Prieto et Nieto, 2014 ; Schaefer, 2008 ; Upshur et Turner, 1999 ; Weigle, 1998 ; Wigglesworth, 1993 ; Wind et Engelhard Jr., 2013 ; Winke, Gass et Myford, 2011 ; Wiseman, 2012 ; Wolfe *et al.*, 2007 ; Zhang, 2016).

De ce lot, 23 se sont déroulées en expression écrite, 14 en expression orale et 2 à la fois en expression écrite et orale. Dans 35 études sur 39, l'écart de sévérité entre l'examinateur le plus sévère et l'examinateur le plus clément était d'au moins 1 logit. Pour l'ensemble des 39 études, l'écart moyen, au sein de la même étude, entre l'examinateur le plus clément et le plus sévère était de 2,60 logits avec un écart type de 1,36, ce qui montre une variabilité importante entre les niveaux de sévérité des

examinateurs des diverses études[8]. Cet écart de 2,60 logits peut être illustré par la situation hypothétique suivante. Supposons deux examinateurs évaluant la même performance avec la même grille d'évaluation. Ceux-ci hésitent entre accorder un « A » ou un « B » à cette performance. L'un des examinateurs est plus sévère que son collègue, son niveau de sévérité étant plus élevé de 2,60 logits. La probabilité que l'examinateur le plus sévère accorde un « A » (P = 0,5), la note la plus élevée, sera 0,54 fois plus faible que la probabilité que l'examinateur le plus clément accorde un « A » (P = 0,93). Une autre manière de présenter l'importance de cet écart est que, pour les 28 études pour lesquelles l'information est disponible, l'étendue des niveaux de sévérité des examinateurs (2,60 logits) représente 21 % de l'étendue des niveaux d'habileté des candidats évalués (12,30 logits). Un tel écart entre deux examinateurs n'est pas trivial et son impact sur les notes accordées ne peut être ignoré. Pourtant, certaines études montrent des écarts de 3, 4, voire 6 logits (respectivement : Casanova et Demeuse, 2011 ; Eckes, 2005 ; Korenovska, 2013) entre les examinateurs les plus cléments et sévères, et ce bien que ces études aient été faites avec des examinateurs expérimentés, travaillant pour de grandes organisations de testing[9] encadrant rigoureusement le travail de leurs examinateurs. Il ne semble pas y avoir de différence importante entre les examinateurs en expression orale et écrite, les écarts de niveaux de sévérité étant aussi importants dans les deux cas. Les 23 recherches en expression écrite ont un écart moyen de 2,30 logits, un écart type de 1,5 et les 14 recherches en expression orale ont un écart moyen de 3,05 logits et un écart type de 1,09. Les écarts entre les niveaux de sévérité des examinateurs représentent donc un problème important pour les examens à forts enjeux et, dans bien des cas, le résultat d'un candidat dépend en partie du niveau de sévérité des examinateurs.

---

[8] *Stricto sensu*, les écarts en logit entre des études utilisant des échelles de notation ayant un nombre différent d'échelons ne peuvent être directement comparés, car le nombre d'échelons influence la valeur de l'étendue des mesures obtenues en logit (c'est l'un des paramètres du modèle, voir l'équation 4 ci-dessus).

[9] Chambre de Commerce et d'Industrie de Paris (CCIP), TestDaF Institut, Educational Testing Service (ETS).

2.3.1 Facteurs associés au niveau de sévérité

Puisque les écarts de niveaux de sévérité entre examinateurs sont importants, plusieurs études ont cherché à identifier les facteurs expliquant ces écarts. Elles ont, pour cela, retenu trois types de facteurs potentiels : langagiers, formationnels et expérientiels.

2.3.2 Facteurs langagiers

Beaucoup d'études ont cherché à savoir s'il y avait une différence de niveau de sévérité des examinateurs selon la langue maternelle de ceux-ci et leurs compétences langagières. Faites dans le domaine de l'évaluation en langue étrangère, ces recherches arrivent toutes à la conclusion qu'il n'y a pas de lien entre la langue maternelle des examinateurs et leur niveau de sévérité. En d'autres termes, les examinateurs locuteurs natifs ne sont pas plus ou moins sévères que les locuteurs non natifs lorsqu'ils évaluent une performance en cette langue (Brooks, 2013 ; Brown, 1995 ; Huang, Alegre et Eisenberg, 2016 ; Huang et Juhn, 2015 ; Johnson et Lim, 2009 ; Wei et Llosa, 2015 ; Zhang et Elder, 2011). Quant aux compétences langagières des examinateurs, opérationnalisées par le diplôme obtenu par ceux-ci, il ne semble pas y avoir de lien entre le niveau de sévérité d'un examinateur évaluant une performance dans une langue et son niveau de compétence en cette langue. En d'autres termes, les examinateurs ayant une maîtrise ne sont pas plus sévères que ceux ayant un baccalauréat. Quelques études ont toutefois révélé une exception à ce constat, soit le fait que les examinateurs familiers avec un accent ou compétents en une langue sont plus cléments lorsqu'ils évaluent la phonologie des candidats locuteurs natifs de cette langue (Carey, Mannell et Dunn, 2011 ; Hsieh, 2011 ; Huang et Juhn, 2015 ; Kang, 2008 ; Winke *et al.*, 2011). Bien sûr, ce résultat ne vaut que pour les examinateurs en expression orale.

2.3.3 Facteurs formationnels

Les détails techniques des procédures de formation des examinateurs utilisées par les organisations responsables des examens à forts enjeux relèvent du secret professionnel et ne sont pas publics. Les informations publiées permettent toutefois de dresser un portrait partiel de ces procédures (Elder *et al.*, 2007 ; Elder, Knoch, Barkhuizen et Von Randow, 2005 ; Lim, 2009 ; Shaw, 2002 ; Xi et Mollaun, 2009). En plus des études susmentionnées, la plupart des organisations responsables d'examens à forts enjeux publient sur leur site Web des informations à ce sujet, comme IELTS (https://www.ielts.org/teaching-and-research/examiner-recruitment-and-training) ou le Centre International d'Études Pédagogiques (http://www.ciep.fr/habilitations). Généralement, les examinateurs potentiels pour un test évaluant une langue *X* sont recrutés en fonction de leur formation universitaire et de leur expérience de travail en enseignement de la langue *X*. Une formation de 1er ou de 2e cycle pertinente est demandée (linguistique, linguistique appliquée, enseignement de *X*, langue étrangère), de même qu'une expérience de travail minimale de quelques mois ou années. Les examinateurs potentiels sont ensuite familiarisés avec les aspects théoriques de leur travail : modèle du construit évalué, échelle de niveaux de compétence, fondements théoriques de l'examen, tâches utilisées pour l'examen, critères d'évaluation, aspects logistiques… Les examinateurs potentiels s'entraînent ensuite à évaluer des performances calibrées et ils analysent leurs jugements et les justifient. Il y a généralement, à la fin, un processus formel de calibration et de sélection où les examinateurs potentiels doivent évaluer plusieurs performances et accorder des notes se situant à l'intérieur d'un intervalle de confiance prédéfini. En sus de cette formation initiale, les évaluations des examinateurs en exercice sont généralement surveillées et ces examinateurs suivent des formations continues (McClellan, 2010).

En dépit de la rigueur des processus de formation initiale, les écarts de niveau de sévérité d'examinateurs en exercice relevés à la section 2.3 montrent que l'impact de la formation initiale sur le niveau de sévérité est minime ou alors fugace, puisque la plupart de ces études ont été faites avec des examinateurs expérimentés ayant tous suivi une formation initiale et un processus de sélection. Il est d'ailleurs difficile, pour cette raison, d'étudier les liens entre la formation initiale et le niveau de sévérité des examinateurs, car il y a un biais de sélection, les examinateurs recrutés par les recherches ayant tous réussi leur processus de sélection. Certaines études ont toutefois cherché à évaluer l'efficacité d'une formation initiale offerte à des examinateurs débutants en utilisant le devis quasi expérimental suivant : une familiarisation minimale au contexte et aux outils d'évaluation, sans formation rigoureuse, suivie d'une première série d'évaluations, puis une formation en bonne et due forme. Au moins une autre séance d'évaluations suit cette formation. 6 études ont un tel devis, soit 3 en expression écrite (Elder *et al.*, 2007 ; Fahim et Bijani, 2011 ; Weigle, 1998) et 3 en expression orale (Davis, 2016 ; H. J. Kim, 2011 ; Kondo, 2010). Une étude supplémentaire a un devis similaire, où des examinateurs débutants suivent 2 séances de formation pour ensuite faire une session d'évaluation réelle (Lumley et McNamara, 1995).

Les résultats de ces études sont mitigés. Fahim et Bijani ont utilisé le devis décrit ci-dessus, mais en incluant deux séances de formation entre les évaluations préformation et postformation. Pour les 12 examinateurs, avant les formations, l'écart du niveau de sévérité était de 6,90 logits et l'écart type de 2,20. 7 examinateurs avaient un niveau supérieur à 1 ou inférieur à -1 logit. Suite aux deux formations, l'écart n'était plus que de 2,60 logits et l'écart type de 0,84. 6 des 7 examinateurs avaient maintenant un niveau de sévérité compris entre +1 et -1 logit et le seul examinateur ayant encore un niveau à l'extérieur de ces bornes avait tout de même substantiellement recentré son niveau de sévérité, passant de -3,60 à -1,60 logits. Weigle a suivi le même devis, mais

avec une seule séance de formation. Il y avait 16 examinateurs en tout, soit 8 débutants et 8 expérimentés. Les 8 examinateurs débutants avaient, avant la formation, un écart de 1,51 logit et un écart type de 0,43. Parmi les 16 examinateurs, le plus sévère et le plus clément étaient des débutants, et l'écart entre le débutant le plus sévère et l'expérimenté le plus sévère était de 1,01 logit. Après la formation, l'écart pour les 8 examinateurs débutants était réduit à 1,00 logit et l'écart type à 0,38. La différence entre les examinateurs débutants et expérimentés les plus sévères n'était plus que de 0,31 logit. Les deux examinateurs débutants qui étaient les plus sévères et cléments étaient, après la formation, dans la moyenne. Dans ces deux cas (Fahim et Bijani ; Weigle), la formation a été un succès et les niveaux de sévérité des examinateurs étaient plus acceptables après la formation qu'avant.

Ce n'est pas le cas pour les 4 autres études recensées, où les effets de la formation ont été neutres (Davis, 2016 ; H. J. Kim, 2011 ; Kondo, 2010), voire négatifs (Elder *et al.*, 2007). Dans les 3 premiers cas, le même phénomène s'est produit. Suite à la première formation (l'étude de H. J. Kim a 3 temps de mesure et 2 séances de formation), un examinateur débutant qui avait un niveau de sévérité trop éloigné de celui de ses pairs voit son niveau se recentrer, mais un examinateur débutant ayant un niveau acceptable au premier temps voit son niveau devenir trop élevé (Davis) ou faible (H. J. Kim ; Kondo) suite à la formation. L'étude de Davis avait également deux examinateurs qui, suite à la formation, ont vu leur niveau de sévérité s'améliorer pour ensuite empirer lors des deux temps de mesure subséquents. L'étude d'Elder *et al.*, elle, avait 2 examinateurs débutants. Le premier est passé de trop clément à trop sévère, tandis que le second est passé d'acceptable à légèrement sévère. Finalement, l'étude de Lumley et McNamara avait 3 temps de mesure : 13 examinateurs étaient présents aux deux premiers et 4 au dernier. Lors du premier temps de mesure, 2 examinateurs sur 13 avaient un niveau de sévérité inadéquat. Suite à la deuxième formation, l'un de ces examinateurs avait un niveau acceptable et l'autre avait un

niveau toujours trop élevé. Par ailleurs, un examinateur ayant un niveau acceptable au premier temps était maintenant trop clément. Au troisième temps, 3 des 4 examinateurs avaient un niveau inadéquat, dont l'examinateur qui avait sensiblement amélioré son niveau entre le premier et le deuxième temps de mesure.

Considérant ces 7 études, il est loin d'être clair que les formations offertes ont été la cause de l'amélioration du niveau de sévérité des examinateurs. D'une part, le fait que certains examinateurs aient vu leur niveau de sévérité s'éloigner davantage de celui de leurs collègues après la formation remet en question l'efficacité de cette dernière. Il faut toutefois préciser qu'il est rare qu'un examinateur trop sévère ou clément le devienne davantage suite à une formation, bien que cela survienne (par exemple : l'examinateur 1 de l'étude de Kondo ou celui de l'étude de Lumley et McNamara). Généralement, le problème est que, suite à une formation, un examinateur ayant un niveau de sévérité moyen devient trop sévère ou clément, sans que ce changement ne soit nécessairement attribuable à la formation. D'autre part, deux études de Lim (2009, 2011) offrent un contre-exemple instructif. Dans les deux études, il y avait des examinateurs débutants ayant suivi avec succès une formation initiale et dont les débuts professionnels ont été suivis par le chercheur, sans que ces débutants ne reçoivent de formation spécifique supplémentaire visant à contrôler leur niveau de sévérité. La première étude avait 4 examinateurs débutants, dont 2 avaient un niveau de sévérité acceptable pour l'ensemble de la période étudiée. Des 2 autres, l'un était trop clément et l'autre trop sévère au premier temps de mesure, mais leur niveau s'est graduellement recentré et, à partir du troisième temps de mesure, leur niveau de sévérité se confondait à celui de leurs collègues. La deuxième recherche comprenait, au total, 6 examinateurs débutants. 3 de ces examinateurs avaient un niveau de sévérité problématique au premier temps de mesure et ce niveau s'est recentré de manière à ce que, à compter du troisième temps de mesure, leur niveau de sévérité était semblable au niveau de leurs collègues. Il est donc difficile de

départager ce qui revient à la formation de ce qui revient à l'expérience professionnelle.

2.3.4 Facteurs expérientiels

En accord avec les résultats présentés à la section précédente, les études ayant comparé les niveaux de sévérité d'examinateurs débutants et expérimentés n'ont généralement pas trouvé de différences significatives entre ces deux groupes (Attali, 2016 ; Hsieh, 2011 ; H. J. Kim, 2011 ; Myford, Marr et Linacre, 1996). Cela est logique, puisque les études de la section précédente ont montré que, dans certains cas, les examinateurs débutants ayant un niveau de sévérité trop élevé ou bas peuvent, après une formation ou avec le temps, modifier leur niveau de sévérité afin qu'il se rapproche du niveau de leurs collègues. Cela suppose qu'il n'y a pas de différence importante entre les niveaux de sévérité des examinateurs débutants et plus expérimentés. Une seule étude a observé un lien entre le niveau de sévérité et l'expérience, les examinateurs débutants de cette étude étant plus sévères que leurs collègues plus expérimentés (Huang et Juhn, 2015), mais cette étude ne concernait que l'évaluation de la phonologie. Tel qu'indiqué à la section 2.3.2, les examinateurs plus familiers avec les accents des candidats étaient plus cléments avec ceux-ci, l'expérience tenant lieu de familiarité avec un accent. Il semble par conséquent qu'il n'y a pas de lien entre l'expérience et le niveau de sévérité des examinateurs.

2.3.5 Synthèse des facteurs de la sévérité

Il n'y a donc, hors le cas précis de l'évaluation des critères phonologiques en expression orale, aucun facteur du niveau de sévérité des examinateurs identifié avec certitude par la recherche. Les résultats ci-dessus sur le niveau de sévérité ne révèlent rien sur la stabilité temporelle de celui-ci. L'indépendance entre le niveau de sévérité des examinateurs, leur expérience et leur formation est compatible avec un niveau de sévérité temporellement stable ou, au contraire, instable et changeant aléatoirement

d'une session de notation à une autre, et ce parce que la stabilité temporelle du niveau de sévérité s'étudie de manière intraindividuelle, tandis que l'association entre le niveau de sévérité et d'autres variables est interindividuelle. La possible instabilité temporelle du niveau de sévérité, que Wilson et Case ont nommé « dérive » du niveau de sévérité (1997), pourrait ainsi expliquer l'absence d'explication quant aux facteurs du niveau de sévérité et à l'absence relative d'efficacité des formations faites pour contrôler le niveau de sévérité des examinateurs.

2.4 Dérive temporelle de la sévérité

Huit études en évaluation en langue ont des résultats pertinents sur l'évolution temporelle du niveau de sévérité des examinateurs. De ces 8 études, 5 ont étudié la sévérité d'examinateurs en expression écrite et 3 en expression orale. La dérive temporelle sera définie, dans la section suivante, comme tout écart intraindividuel d'au moins 1 logit.

2.4.1 Expression écrite

Congdon et McQueen (2000) ont étudié le niveau de sévérité de 16 examinateurs évaluant des productions écrites de 8 285 élèves du primaire, en Australie. 12 des 16 examinateurs avaient de l'expérience et 4 étaient débutants. Le niveau de sévérité des examinateurs a été mesuré à 6 reprises sur une période de 8 jours, à l'aide du modèle de Rasch à multifacettes. Les résultats montrent une dérive temporelle importante du niveau de sévérité de 7 examinateurs, ceux-ci ayant un écart d'au moins 1 logit entre leur niveau de sévérité le plus bas et le plus élevé lors des 6 temps de mesure. L'écart moyen entre le niveau de sévérité le plus élevé et le plus bas, au cours des 6 temps de mesure, pour l'ensemble des 16 examinateurs, est de 0,98 logit avec un écart type de 0,43 et seuls 2 examinateurs ont un écart inférieur à 0,50 logit pour les 6 temps de mesure.

Wolfe *et al.* (2007) ont étudié le niveau de sévérité de 101 examinateurs expérimentés, évaluant des productions écrites de 51 233 élèves de 12^{e} année d'écoles secondaires américaines. Ils ont, à l'aide du modèle de Rasch à multifacettes et d'indices développés spécifiquement pour ce faire, mesuré le niveau de sévérité des examinateurs à 8 reprises sur une période de 4 jours, mais en comparant le niveau de sévérité à la mesure de référence obtenue lors du premier temps de mesure. Les résultats individuels ne sont pas disponibles, mais, selon les indices développés par les auteurs, 33 % des examinateurs ont un niveau de sévérité statistiquement différent à au moins un des 7 temps de mesure par rapport à la mesure de référence au premier temps (au seuil α de 0,05). Dans presque tous les cas, la dérive observée était graduelle, seuls 8 % des examinateurs ayant un niveau de sévérité changeant brusquement entre deux temps de mesure.

Lim (2009), lui, a étudié l'évolution du niveau de sévérité de 10 examinateurs évaluant les productions écrites d'anglais langue seconde d'un nombre non spécifié de candidats, mais clairement supérieur à 1 000. Quatre des 10 examinateurs étaient nouveaux, les autres ayant au moins plusieurs mois d'expérience au moment de la collecte de données. Les données ont été collectées lors de deux périodes différentes, la première ayant 7 temps de mesure sur 21 mois et la seconde, 5 temps de mesure sur 15 mois et elles ont été analysées à l'aide du modèle de Rasch à multifacettes. Les résultats de la première période montrent une dérive temporelle importante de plus de 1 logit pour 2 des examinateurs, l'écart moyen étant de 0,71 logit et l'écart type de 0,44. Sur les 7 examinateurs, seuls 2 ont un écart maximal inférieur à 0,50 logit. En revanche, les résultats de la deuxième période révèlent une absence de dérive temporelle du niveau de sévérité, l'écart moyen étant de 0,30 logit et l'écart type de 0,11, un seul examinateur ayant un écart supérieur à 0,50 logit, son écart étant de 0,51 logit. Lim (2011) a utilisé les données de sa thèse, présentées ci-dessus, mais les a analysées différemment. 11 examinateurs ont été suivis sur 3 périodes de temps

distinctes, de 12, 21 et 13 mois, avec un temps de mesure à chaque mois. Bien que les données aient été analysées avec le modèle de Rasch à multifacettes, les résultats ne sont présentés que sous forme graphique et l'unité de mesure n'est pas le logit, mais bien l'unité de l'échelle d'appréciation d'origine, soit une échelle à 9 valeurs. Il est donc difficile d'interpréter les graphiques, visuellement chargés, mais il semble que 5 des 10 examinateurs aient un écart supérieur à 0,5 point entre leur niveau de sévérité maximal et minimal, ce qui représente une dérive temporelle non négligeable pour ces examinateurs.

Leckie et Baird (2011) ont étudié 689 examinateurs lors de l'évaluation de productions écrites d'élèves anglais de 14 ans. De ces 689 examinateurs, 135 étaient chef d'équipe, 372 avaient au moins un an d'expérience et 182 étaient débutants. Les notes accordées à 84 productions écrites étalons ont été analysées à l'aide d'un modèle linéaire généralisé multiniveaux, et les données ont été collectées à 5 temps de mesure distincts. Le texte ne précise toutefois pas la durée totale de la collecte, mais il est vraisemblable que le tout ait eu lieu en quelques jours, voire 1 ou 2 semaines au maximum. Les résultats ne sont présentés que sous forme graphique, et ce pour un sous-échantillon des examinateurs étudiés. Il semble qu'il y a dérive temporelle du niveau de sévérité de bien des examinateurs, qu'ils soient ou non expérimentés, certains écarts étant de 0,75 pour une échelle d'appréciation à 9 valeurs, ce qui est substantiel. La présence de dérive temporelle est également confirmée par le fait qu'un modèle d'analyse permettant au paramètre de sévérité des examinateurs de varier en fonction du temps de mesure était significativement mieux ajusté aux données qu'un modèle forçant le niveau de sévérité des examinateurs a être stable pour les 5 temps de mesure.

2.4.2 Expression orale

Lumley et McNamara (1995) ont étudié le niveau de sévérité de 13 examinateurs évaluant l'expression orale d'anglais langue étrangère de professionnels de la santé. Il y a eu 3 temps de mesure sur une période de 20 mois. Des 13 examinateurs, seuls 4 étaient présents aux 3 temps de mesure. Les données, provenant de la performance de 83 candidats, ont été analysées avec le modèle de Rasch à multifacettes. Les résultats indiquent la présence de dérive temporelle du niveau de sévérité chez 3 des 13 examinateurs, ceux-ci ayant un écart supérieur à 1 logit entre leur niveau minimal et maximal. Des 10 autres examinateurs, 9 ont un écart inférieur à 0,50 logit.

H. J. Kim (2011) a, elle, étudié le niveau de sévérité de 9 examinateurs à 3 temps de mesure séparés d'un mois chacun. Les données proviennent d'un test d'expression orale d'anglais langue étrangère et les 9 examinateurs ont évalué 6 candidats à chaque temps de mesure. Les analyses, faites avec le modèle de Rasch à multifacettes, montrent une dérive temporelle importante pour 5 des 9 examinateurs, ceux-ci ayant un écart entre leur niveau de sévérité minimal et maximal supérieur à 1 logit. Pour l'ensemble des 9 examinateurs, l'écart moyen est de 1,09 logit et l'écart type de 0,58. Parmi les 9 examinateurs, une seule a un écart inférieur à 0,50 logit.

Davis (2016) a étudié le niveau de sévérité de 20 examinateurs débutants à 4 temps de mesure sur une période de 4 semaines. Les données proviennent de 240 candidats ayant fait l'examen d'expression orale du TOEFL iBT et elles ont été analysées avec le modèle de Rasch à multifacettes. Les résultats révèlent la présence de dérive temporelle importante pour 5 des 20 examinateurs, ceux-ci ayant un écart supérieur à 1 logit entre leur niveau de sévérité minimal et maximal. Par ailleurs, seuls 3 des 20 examinateurs ont un écart inférieur à 0,50 logit, l'écart moyen étant de 0,86 logit et l'écart type de 0,49.

Mis en parallèle, les résultats de ces 8 recherches sont très intéressants. Pour l'ensemble de celles-ci, une proportion substantielle d'examinateurs voit son niveau de sévérité fluctuer en fonction du temps, soit environ 33 % des examinateurs lorsque les résultats sont cumulés. Cette proportion semble stable, du moins pour les études pour lesquelles les résultats individuels exacts sont disponibles. Elle va de 20 % (Lim, 2009) à 55 % (H. J. Kim, 2011), tout en gardant à l'esprit que le nombre total d'examinateurs pour certaines études est faible et qu'il faut être prudent avec ces pourcentages. Les deux études présentant uniquement leurs résultats sous forme graphique (Leckie et Baird, 2011 ; Lim, 2011) semblent en accord avec ces proportions. Malheureusement, rien dans les résultats de ces études ne permet d'identifier les examinateurs dont le niveau de sévérité dérive. Si certaines études (Lim, 2009 ; H. J. Kim, 2011) laissent croire que les examinateurs débutants sont plus à risque, les études de Wolfe *et al.* et de Leckie et Baird montrent bien que les examinateurs expérimentés ont aussi des problèmes de dérive temporelle. Aucune autre variable ne semble liée à la dérive temporelle dans ces études. Mais en dépit du grand intérêt des résultats de ces études, une limite importante demeure. Presque toutes ces études comptent peu de temps de mesure (Davis, 2016 ; H. J. Kim, 2011 ; Leckie et Baird, 2011 ; Lumley et McNamara, 1995) ou alors elles sont faites sur une très courte période de temps (Congdon et McQueen, 2000 ; Wolfe *et al.*, 2007). Seules les 2 études de Lim ont un nombre raisonnable de temps de mesure échelonnés sur une période de plusieurs mois consécutifs. Malgré leur grande qualité, ces 2 études ont tout de même des limites. La première étude (2009) concerne deux périodes disjointes, présentant chacune peu de temps de mesure (7 et 5) et peu d'examinateurs (6 et 7). De plus, chaque temps de mesure comprend 3 mois, ce qui peut gommer d'éventuelles dérives temporelles à court ou moyen terme. La deuxième étude a peu d'examinateurs (11) et elle présente ses résultats d'une manière difficile à interpréter, car strictement graphique. Qui plus est, les données de cette étude semblent être les mêmes que les données de l'étude précédente (2009) et il est difficile de savoir si les examinateurs présentés dans les deux études diffèrent ou non.

Finalement, toutes les études recensées ici ont choisi de modéliser le niveau de sévérité des examinateurs en fonction de périodes de temps égales, et non en fonction du nombre de candidats évalués. Cela ne pose pas problème si les examinateurs évaluent le même nombre de candidats pour chaque période de temps, mais ce n'est pas le cas dans la plupart des études recensées. Par exemple, dans l'étude de Lim (2011), un examinateur évalue 419 candidats lors de ses 3 premiers mois d'activité et 1 307 candidats lors des 3 mois suivants. Or, au-delà du temps qui passe, il est probable que ce soit la charge de travail qui influence le niveau de sévérité des examinateurs, si influence il y a. Les résultats de ces 8 recherches ne permettent donc pas de conclure quant à l'importance de la dérive temporelle et à sa constance. Il est possible que les dérives temporelles observées dans ces études soient le résultat de niveaux de sévérité exceptionnellement élevés ou bas à un seul temps de mesure et que, si le suivi avait été plus long, cette valeur exceptionnelle aurait apparu comme extrême et négligeable. Le faible nombre de temps de mesure empêche aussi la détection de tout effet potentiel de variations saisonnières et cycliques, ce qui diminue grandement l'utilité d'avoir recours à des mesures répétées, seule la tendance générale pouvant être détectée.

2.5 Synthèse générale sur le niveau de sévérité

Une synthèse de cette revue de littérature fait ressortir quelques faits importants. Qu'ils soient expérimentés ou non, qu'ils travaillent pour une organisation encadrant rigoureusement leurs examinateurs ou non, une proportion non négligeable d'examinateurs ont un niveau de sévérité suffisamment bas ou élevé par rapport à leurs collègues pour que l'identité de l'examinateur ait un impact sur les notes reçues par une performance. Le problème est que, indépendamment du niveau d'habileté d'un candidat, si les examinateurs d'une épreuve ont des niveaux de sévérité trop éloignés, le fait que l'examinateur A évalue une performance, plutôt que l'examinateur B, aura un impact important sur les notes accordées. Les études

transversales montrent que, à n'importe quel temps de mesure, le niveau de sévérité de plusieurs examinateurs est très élevé ou très faible par rapport au niveau de sévérité de leurs collègues. Plus préoccupant encore, les études longitudinales révèlent que, même lorsqu'il n'y a aussi peu que 3 temps de mesure, le niveau de sévérité de certains examinateurs dérive significativement, si bien qu'environ un tiers des examinateurs étudiés ont un niveau de sévérité fluctuant de plus ou moins 1 logit, ou plus, au cours de la période étudiée.

Les formations faites pour rendre acceptable le niveau de sévérité d'examinateurs donnent des résultats mitigés, contradictoires, pour autant que l'on puisse affirmer que ce sont bien les formations qui agissent sur le niveau de sévérité et non seulement l'expérience professionnelle ou le temps qui passe. Il se peut que la dérive temporelle du niveau de sévérité explique l'effet de la formation, c'est-à-dire que le changement de niveau de sévérité suite à une formation soit une coïncidence et ne soit qu'une dérive du niveau de sévérité. Cela pourrait expliquer pourquoi, suite à une formation, certains examinateurs voient leur niveau de sévérité devenir plus extrême, alors que d'autres voient leur niveau s'améliorer pour ensuite revenir à un niveau trop élevé ou faible. Il pourrait également simplement s'agir, dans certains cas, d'une régression à la moyenne. Les effets respectifs de la formation et de l'expérience sont d'autant plus inextricables que les deux effets sont sociaux : puisque les notes accordées par les examinateurs ne peuvent être comparées à un ensemble de notes « objectivement justes et exactes », la rétroaction qu'un examinateur obtient vient essentiellement de la comparaison entre les notes qu'il accorde et les notes que ses collègues et superviseurs accordent. Le développement professionnel d'un examinateur repose en partie sur cette confrontation entre son jugement et celui de ses pairs et la qualité de cette dynamique sociale pourrait avoir un effet propre sur l'évolution temporelle du niveau de sévérité (Crimmins, Nash, Oprescu, Alla, Brock, Hickson-Jamieson et

Noakes, 2016 ; Klenowski et Wyatt-Smith, 2014). Il se pourrait donc que les examinateurs aient un impact sur les niveaux de sévérité de leurs collègues.

Les études recensées ne permettent pas de trancher ces questions avec certitude, il faudrait pour cela en savoir davantage sur la dérive temporelle du niveau de sévérité d'examinateurs, particulièrement en ce qui concerne l'évolution du niveau de sévérité sur un très grand nombre de temps de mesure (p. ex. 30) répartis sur plusieurs années consécutives. Il faudrait également modéliser l'évolution du niveau de sévérité en fonction du nombre de candidats évalués et non seulement en fonction de périodes de temps équidistantes. Il serait aussi pertinent de pouvoir approfondir les connaissances fournies par les 2 études de Lim (2009, 2011) sur l'évolution temporelle du niveau de sévérité des examinateurs débutants, afin de savoir si le niveau de sévérité de ces derniers est plus à risque de dérive temporelle. Pour terminer, étudier davantage la dynamique sociale d'une éventuelle influence des examinateurs sur le niveau de sévérité de leurs collègues pourrait nous permettre de mieux comprendre cet aspect important du travail des examinateurs.

2.6 Objectifs spécifiques de recherche

En vertu de quoi, cette recherche a les 3 objectifs spécifiques suivants :

1 – Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du nombre de candidats évalués et du temps chronologique;

2 – Comparer l'évolution du niveau de sévérité des examinateurs débutants et expérimentés;

3 – Comparer l'évolution du niveau de sévérité d'examinateurs travaillant ensemble.

CHAPITRE III

MÉTHODOLOGIE

Pour atteindre les 3 objectifs spécifiques de recherche énumérés ci-dessus, un devis rétrospectif longitudinal sera utilisé. Les données de cette thèse sont des données secondaires (Hox et Boeije, 2005) provenant de tests de français langue étrangère que des candidats ont faits d'octobre 2010 à avril 2014, dans un centre de passation situé au Québec. Les composantes méthodologiques de cette thèse sont expliquées dans les pages qui suivent. Premièrement, l'échantillon sera caractérisé. Deuxièmement, les instruments utilisés pour collecter les données seront présentés, de même que le déroulement du test lui-même. Quelques informations sur le processus de collecte et de transmission des données suivront, puis les techniques analytiques seront détaillées en profondeur, de même que les logiciels qui seront utilisés pour faire ces analyses. Les considérations éthiques et le calendrier de réalisation termineront ce chapitre. Il est à noter que la section « Méthodes d'analyse des données » est particulièrement longue, mais cette longueur est due à la nature de ce projet de thèse, où certains choix analytiques doivent être pleinement justifiés et expliqués.

3.1 Participants

3.1.1 Candidats

Les données proviennent de 3 333 évaluations faites pour l'épreuve d'expression orale de l'une des variantes du TEF (TEF, TEF Naturalisation, TEFaQ) entre octobre 2010 et avril 2014, dans un même centre de test officiel situé au Québec. Aucune information sociodémographique n'est disponible pour les candidats. Les 3 333

évaluations ne représentent pas 3 333 candidats différents, puisque certains candidats ont fait plusieurs fois l'épreuve d'expression orale au cours de cette période. Le nombre exact de candidats ayant fait l'épreuve au moins deux fois est inconnu, ainsi que, par conséquent, le nombre exact de candidats distincts parmi les 3 333 évaluations. L'absence de variables sociodémographiques fait en sorte que les candidats ayant fait l'examen plus d'une fois ne peuvent être identifiés. Ce nombre est toutefois relativement faible, selon les estimations obtenues du centre de test concerné, inférieur à 100. Techniquement, le fait que certaines des évaluations aient été faites par les mêmes candidats introduit un problème de dépendance dans les données, ce qui peut biaiser les estimés de niveau d'habileté des candidats. Ce problème est néanmoins négligeable dans la situation actuelle, où le paramètre d'intérêt pour cette thèse est le niveau de sévérité des examinateurs. Le modèle de mesure Rasch à multifacettes suppose une inévitable violation de l'indépendance locale, puisque les notes accordées à un même candidat par deux examinateurs différents ne sont pas stochastiquement indépendantes. Le modèle est toutefois robuste à cette violation et, dans un contexte typique, l'effet sur les estimés des niveaux de sévérité des examinateurs est négligeable (de Jong et Linacre, 1993 ; Linacre, 2009).

3.1.2 Examinateurs

Un total de 20 examinateurs, soit 6 hommes et 14 femmes, ont évalué au moins un candidat pour les 3 333 évaluations faites durant cette période. Tous les examinateurs sauf 2 avaient le français comme langue maternelle et de scolarité. Les 2 exceptions avaient le roumain comme langue maternelle et le français comme langue de scolarité postsecondaire. Tous ces examinateurs avaient de l'expérience professionnelle comme enseignant de français langue étrangère avant de travailler comme examinateur pour le TEF. De ces 20 examinateurs, 3 travaillaient déjà comme examinateur en octobre 2010 et avaient donc de l'expérience comme examinateur.

Parmi les 17 autres examinateurs, 15 n'avaient aucune expérience comme examinateur pour le TEF avant d'être embauchés à ce titre. Ces 15 examinateurs ont donc été formés à l'interne, par le responsable TEF du centre de test. Cette formation maison avait une partie théorique, comprenant une familiarisation avec les concepts évalués, le test lui-même, les niveaux de compétence du Cadre européen commun de référence pour les langues et la grille d'évaluation. La partie pratique de la formation consistait en l'observation d'authentiques séances de test et en l'évaluation simulée de performance. Finalement, les examinateurs débutants étaient toujours jumelés à un examinateur expérimenté agissant à titre de mentor. Deux examinatrices ont été embauchées au centre de test alors qu'elles avaient déjà de l'expérience comme examinatrices du TEF. Puisque le responsable TEF du centre est demeuré le même durant toute la période étudiée, cela signifie que 17 examinateurs sur 20 ont été formés par la même personne, essentiellement de la même manière – puisque les 3 examinateurs travaillant en octobre 2010 sont le responsable TEF lui-même et deux examinatrices ayant été formées par ce dernier. Les 3 autres examinateurs, soient le responsable TEF du centre et les 2 examinatrices ayant une expérience préalable, ont été formés et certifiés par la CCIP, qui est l'organisme responsable du TEF.

Au cours de la période étudiée, les 20 examinateurs ont évalué de 19 à 1 196 candidats (m = 333 ; s = 386) et ont travaillé de 3 à 37 mois (m = 13 ; s = 8,7), pas nécessairement consécutifs puisque certains examinateurs ont pris des pauses pour diverses raisons durant cette période. Le tableau 3.1 présente les statistiques individuelles des examinateurs. Le code de l'examinateur inclut la lettre « H » ou « F » pour indiquer le genre de l'examinateur.

Tableau 3.1
Liste des examinateurs

Examinateur	Nombre de mois actifs	Plus longue période de travail mensuel consécutif	Nombre de candidats évalués
1H	29	29	591
2F	10	5	45
3F	19	8	56
4F*	41	18	1 175
5F*	37	29	1 196
6F*	26	5	152
7F*	19	19	480
8F*	12	12	273
9H*	9	2	19
10F*	3	3	27
11H*	8	8	95
12F	18	9	376
13H*	17	17	1 076
14F*	6	6	30
15F*	10	10	200
16H*	11	4	64
17F*	9	6	241
18F*	8	8	120
19H*	7	5	89
20F	6	6	361

* L'examinateur était débutant lors de sa première évaluation au centre de test

3.2 Instrument

3.2.1 Test

Toutes les données proviennent de l'épreuve d'expression orale du TEF ou de l'une de ses variantes, un test visant à décrire le niveau de compétence atteint par un candidat. À l'époque où les données ont été collectées, il existait trois variantes du test : le TEF, le TEFaQ et le TEF Naturalisation. Dans les trois cas, l'épreuve d'expression orale était identique et seuls les thèmes utilisés comme mise en situation différaient en partie d'une variante à l'autre. Voici la description de l'épreuve d'expression orale. Cette épreuve est constituée de deux tâches d'évaluation et le

candidat est évalué par deux examinateurs. La première tâche consiste en un jeu de rôle où le candidat doit simuler un appel téléphonique afin d'obtenir de l'information à propos d'un produit ou d'un service offert dans une publicité qu'il a lue. La publicité est fournie au candidat et celui-ci a 10 minutes pour se préparer. La conversation doit durer environ 5 minutes et un premier examinateur joue le rôle de l'interlocuteur, soit la personne offrant le produit ou le service offert dans la publicité. Le candidat doit poser un maximum de questions pertinentes, tout en s'assurant de bien comprendre les réponses fournies par son interlocuteur afin d'orienter son questionnement. L'autre examinateur est à l'écart, il chronomètre et n'intervient pas dans la conversation. La seconde tâche est un autre jeu de rôle au cours duquel le candidat doit convaincre son ami, joué par le deuxième examinateur, de faire une activité quelconque, décrite sur un document qui est remis au candidat pour qu'il puisse se préparer. La préparation dure 10 minutes et la conversation elle-même doit durer environ 10 minutes. Le candidat doit présenter l'information fournie par le document sur l'activité et il doit ensuite mobiliser le plus d'arguments convaincants possible, tout en répondant aux objections soulevées par son interlocuteur. L'autre examinateur, qui jouait l'interlocuteur lors de la première tâche, se tient à l'écart et chronomètre la discussion. À la fin de l'épreuve, le candidat quitte la salle et les deux examinateurs évaluent indépendamment la performance du candidat et, lorsqu'ils ont fini l'évaluation individuelle, ils procèdent à l'arbitrage. Les deux évaluations sont mises en commun et les deux examinateurs doivent se mettre d'accord sur l'évaluation finale accordée à la performance du candidat. Cet arbitrage peut prendre la forme d'une simple moyenne entre les notes des deux examinateurs, du choix de la note de l'un des examinateurs au détriment de la note accordée par l'autre examinateur, ou alors d'un compromis résultant en une note située entre les deux notes des examinateurs. Les évaluations individuelles sont conservées telles quelles et l'évaluation finale est ajoutée sur une troisième grille d'évaluation. Les données de cette étude sont les données des évaluations individuelles, faites par les examinateurs de manière indépendante et sans consultation.

3.2.2 Grille d'évaluation

La grille d'évaluation utilisée ne peut être présentée en détail pour des raisons de propriété intellectuelle et de sécurité du test. Il s'agit d'une grille analytique descriptive à 12 critères d'évaluation individuellement notés. Il y a 3 critères communicatifs pour évaluer la première tâche, 3 critères pour la seconde et 6 critères linguistiques pour évaluer la grammaire, le lexique et la phonologie de l'ensemble de la performance du candidat. Les 12 critères utilisent la même échelle d'appréciation à 21 valeurs, allant de 0 à 20. Ces valeurs sont arrimées aux 6 niveaux de compétence du *Cadre européen commun de référence pour les langues* (CECRL ; Conseil de l'Europe, 2005), auxquels est ajouté un niveau inférieur pour les candidats n'ayant aucune compétence en français. Chaque examinateur assigne une note par critère, pour un total de 12 notes pour chaque performance évaluée, ce qui donne un score total allant de 0 à 240 pour chaque candidat.

3.3 Déroulement de la collecte de données

Les données utilisées dans cette thèse sont des données secondaires. Les données ont originalement été collectées en temps réel pour des motifs de contrôle de la qualité par le responsable TEF du centre de test. Les données étaient saisies dans un fichier Excel après chaque session de test. Les données ont été fournies au chercheur et auteur de cette thèse 3 mois après la fin de la période de collecte, avec la permission des autorités compétentes du centre de test.

3.4 Méthodes d'analyse des données

L'atteinte des objectifs spécifiques de recherche suppose deux étapes distinctes d'analyse des données. Premièrement, estimer le niveau de sévérité des examinateurs à différents intervalles de temps. Deuxièmement, utiliser les estimés obtenus à la première étape et les analyser en tant que variable dépendante à l'aide de différentes

techniques associées à l'étude des séries chronologiques. Voici les détails des deux étapes.

3.4.1 Estimation des niveaux de sévérité des examinateurs

Suite à Casanova et Demeuse (2016), et pour exactement les mêmes raisons, les notes accordées par chaque examinateur seront regroupées en 3 ensembles, en fonction des critères d'évaluation de la grille utilisée. Le premier ensemble sera constitué des 3 notes accordées pour évaluer la première tâche communicative (A), le deuxième ensemble des 3 notes accordées pour la deuxième tâche (B) et le troisième ensemble regroupera les 6 critères linguistiques (L pour « langue »). Un niveau de sévérité distinct sera estimé pour chacun de ces 3 ensembles. Chaque examinateur aura donc 3 estimés différents de niveau de sévérité, ce qui est compatible avec la conception de la sévérité présentée à la section 2.1. Toutefois, puisque la procédure suivie sera la même quel que soit le niveau de sévérité estimé (A, B ou L), la suite de cette section ne tiendra pas compte de cette différence et utilisera le singulier pour désigner le « niveau de sévérité ». Toutes les analyses pour estimer le niveau de sévérité partagent une modélisation commune du modèle de Rasch à multifacettes et les estimations seront faites en 2 étapes.

À la première étape, 4 facettes seront modélisées pour estimer les niveaux de sévérité des examinateurs, soit les candidats, les examinateurs, les critères d'évaluation et le temps. Les paramètres correspondants sont le niveau d'habileté des candidats (θ), le niveau de sévérité des examinateurs (α), la difficulté des critères d'évaluation (β) et le temps (κ). Cette dernière facette est une facette dite « factice[10] » ne contribuant pas à l'estimation des 3 autres paramètres du modèle (Eckes, 2011). Elle ne sert qu'à estimer un niveau de sévérité distinct pour chaque intervalle de temps modélisé à la 2^{e} étape. À ces 4 facettes s'ajoute une 5^{e} facette représentant la difficulté relative de

[10] Traduction libre de *dummy facet*. Voir le glossaire.

chaque catégorie de l'échelle d'appréciation utilisée, puisque les données brutes sont polychotomiques (τ). La version *rating scale* du modèle sera utilisée, car chaque critère correspond à la même échelle sous-jacente des niveaux de compétence du CECRL et les examinateurs sont supposés avoir une compréhension commune de l'échelle. La facette « Examinateurs » sera la facette flottante dont la moyenne sera librement estimée, puisqu'il s'agit de la facette dont les valeurs feront l'objet des analyses subséquentes. Les paramètres des autres facettes seront contraints d'avoir une moyenne de 0 logit. L'équation 5 représente le modèle utilisé à la 1ère étape :

$$\ln\left[\frac{P_{nijx=c}}{P_{nijx=c-1}}\right] = \theta_n - \beta_i - \alpha_j - \kappa_t - \tau_c \qquad (5)$$

Les valeurs estimées à l'étape 1 sont utilisées pour calculer les résidus, soit les différences entre les valeurs attendues par le modèle et les valeurs réellement observées. Ces résidus sont ensuite utilisés, à la 2e étape, comme données pour estimer les valeurs du paramètre d'interaction entre la facette du niveau de sévérité et du temps (α_{jt}), les valeurs des autres paramètres étant ancrées aux valeurs estimées à l'étape 1 (Linacre, 2017a). L'équation 6 représente la 2e étape.

$$\ln\left[\frac{P_{nijx=c}}{P_{nijx=c-1}}\right] = \hat{\theta}_n - \hat{\beta}_i - \hat{\alpha}_j - \hat{\kappa}_t - \alpha_{jt} - \hat{\tau}_c \qquad (6)$$

La différence entre les deux modèles qui seront utilisés à la 2e étape réside en la définition des éléments de la facette « temps » (κ_t). Dans les deux cas, le modèle inclut une interaction entre les facettes de niveau de sévérité des examinateurs et de temps (α_{jt}), ce qui permettra d'obtenir un estimé du niveau de sévérité différent pour chaque combinaison sévérité × temps, tout en assurant la liaison entre les divers éléments de toutes les facettes, ce qui permettra d'estimer conjointement les valeurs de tous ces éléments sur la même échelle en logit. Ce devis a posteriori est un devis en spirale incomplet lié[11], aussi appelé échantillonnage matriciel, où la liaison assure

[11] Voir le glossaire.

un degré de précision satisfaisant, car 3 examinateurs (1H, 4F, 13H) assurent à eux seuls la liaison de toutes les données (Eckes, 2011). Les examinateurs 1H et 4F ont conjointement évalué 145 candidats et les examinateurs 4F et 13H 158 candidats. Les 17 autres examinateurs ont évalué un minimum de 8 candidats avec l'un ou l'autre de ces 3 examinateurs, ce qui assure une liaison complète de toutes les données.

3.4.1.1 Estimation en fonction du nombre de candidats évalués

Dans ce cas, 12 analyses différentes seront faites, soit 1 analyse pour chaque examinateur ayant évalué au moins 100 candidats. Toutes les données sont utilisées, mais les éléments de la facette « Temps » seront définis de manière à ce que chaque intervalle de temps corresponde à exactement 10 candidats évalués par l'examinateur d'intérêt, peu importe le temps chronologique pendant lequel ont eu lieu ces évaluations. Puisque le nombre de candidats évalués par les examinateurs n'est pas nécessairement divisible par 10, le dernier intervalle de temps peut comprendre moins de 10 candidats. Ce choix arbitraire de 10 candidats repose sur deux arguments. Premièrement, le fait d'avoir 10 candidats par intervalle fait en sorte qu'il y a 30 ou 60 notes accordées par chaque examinateur (10 candidats × 3 ou 6 critères), ce qui résulte en des estimés de niveaux de sévérités ayant de petites erreurs types, environ 0,15 logit selon des analyses préliminaires. Deuxièmement, pour une grande partie de la période d'où proviennent les données de cette thèse, les examinateurs évaluaient exactement 5 candidats par jour de travail. Découper le temps en tranches de 10 candidats, soit approximativement 2 jours de travail, est compatible avec un processus de formation continue et de retour sur leur travail fait par les examinateurs.

En revanche, un tel découpage n'est pas assez fin pour étudier spécifiquement l'évolution du niveau de sévérité d'examinateurs débutants, car l'évaluation de 50 candidats ne mènerait ainsi qu'à 5 estimés du niveau de sévérité, ce qui est peu pour tirer quelque conclusion que ce soit. Et, si la littérature n'offre aucune information

concrète sur le processus de transition d'un examinateur débutant à expérimenté et sur le nombre d'évaluations que cela suppose, il semble douteux qu'un examinateur ayant évalué plus de 50 candidats puisse être considéré « débutant », en autant que les candidats évalués par cet examinateur aient présenté une diversité suffisante pour que cet examinateur ait pu développer son jugement. Cela ne signifie pas, non plus, qu'un examinateur ayant évalué 60 candidats soit expérimenté, mais il n'est probablement plus débutant. Par conséquent, pour atteindre le deuxième objectif spécifique, les éléments de la facette « temps » seront également définis afin que chaque intervalle corresponde à l'évaluation de 5 candidats. Ce choix est arbitraire, mais les arguments avancés précédemment tiennent toujours, et les erreurs types valent environ 0,06 logit, ce qui semble acceptable.

3.4.1.2 Estimation en fonction du temps chronologique

Pour l'étude du niveau de sévérité en fonction du temps chronologique, le problème réside en la répartition du travail des examinateurs. Puisque les données sont des données secondaires et que le travail des examinateurs n'était pas organisé en fonction d'une étude subséquente, il est difficile de trouver une période de temps répondant aux critères suivants : pouvoir être découpé en intervalles égaux suffisamment nombreux pour qu'une série chronologique soit possible et inclure au moins deux examinateurs ayant travaillé lors de chacun des intervalles ainsi définis. Il est donc impossible, avec nos données, de faire une seule analyse qui inclurait tous les examinateurs et l'ensemble des données. Les données seront donc analysées séparément afin que chaque analyse respecte les critères susmentionnés. 5 analyses distinctes seront faites et elles incluront, au total, les 9 examinateurs ayant évalué le plus grand nombre de candidats. Le tableau 3.2 donne les informations sur le découpage temporel.

Tableau 3.2
Découpage du temps chronologique

Période	Durée	Fréquence des mesures	*n* mesures	*n* examinateurs	Identité examinateurs
10-2010 à 03-2013	30 mois	1 fois aux 3 mois	10	2	1H; 4F
09-2011 à 02-2013	18 mois	2 fois par mois	34	2	1H; 5F
06-2012 à 11-2013	18 mois	2 fois par mois	36	2	4F; 5F
12-2012 à 09-2013	10 mois	2 fois par mois	19	5	4F 5F; 7F ; 13H; 15F
11-2013 à 04-2014	6 mois	3 fois par mois	18	5	12F; 13H ; 17F; 18 F; 20F

Les données complètes seront utilisées pour chaque analyse, mais les éléments de la facette « temps » seront définis de manière à correspondre aux informations du tableau 3.2 et seuls les estimés des examinateurs concernés seront utilisés dans les séries chronologiques subséquentes. 5 ensembles distincts de résultats seront ainsi produits pour étudier l'évolution du niveau de sévérité en fonction du temps chronologique.

3.4.1.3 Vérification des conditions d'utilisation du modèle de Rasch à multifacettes

Théoriquement, le modèle de Rasch à multifacettes a les mêmes conditions d'utilisation que le modèle élémentaire de Rasch pour données dichotomiques, soit l'unidimensionnalité, l'indépendance locale et la monotonicité de la relation entre le trait latent et la probabilité d'avoir une « bonne » réponse (ou une réponse plus élevée dans le cas de données polychotomiques) (Eckes, 2011 ; Linacre, 2017a). Il est toutefois difficile de tester le respect de deux de ces conditions d'utilisation, et ce pour trois raisons. La première est que, si le modèle utilisé a plus de 2 facettes, les données ne peuvent pas être représentées par une matrice rectangulaire, ce qui empêche toute utilisation des techniques psychométriques habituelles pour étudier la

dimensionnalité d'un ensemble de données, soit l'analyse factorielle ou l'analyse en composantes principales[12]. La deuxième raison est que le modèle à multifacettes est souvent utilisé pour mesurer des performances complexes qui sont conceptuellement multidimensionnelles. Par exemple, dans le cas actuel, il est évident que les habiletés linguistiques et communicationnelles représentent deux dimensions différentes ; il est trivial de former des phrases grammaticales qui sont pourtant vides de sens (Bachman et Palmer, 2010 ; Erickson, Åberg-Bengtsson et Gustafsson, 2015 ; Harding, 2014). La troisième raison est que le modèle de Rasch à multifacettes viole très souvent la condition d'indépendance locale par définition, car le fait que plusieurs examinateurs évaluent les mêmes candidats avec les mêmes critères peut introduire de la dépendance entre les données (de Jong et Linacre, 1993 ; Wilson et Hoskens, 2001). Les notes accordées par deux examinateurs au même candidat ont tendance à être davantage corrélées que les notes accordées par deux examinateurs à deux candidats ayant néanmoins un niveau d'habileté identique (Wang, Su et Qiu, 2014). Or, le modèle suppose que les examinateurs sont des experts indépendants et que les notes de chacun apportent une information pleine et entière, ce qui n'est pas le cas si les examinateurs cherchent à noter pour obtenir un consensus. L'autre menace à l'indépendance locale est le fait, en évaluation des langues, d'utiliser plusieurs critères d'évaluation pour évaluer une même performance. Les notes accordées à des critères similaires sont susceptibles à l'effet de halo, où la note octroyée à un critère dépend en partie de la note octroyée à un critère similaire et non seulement de la performance évaluée (Andrich, Humprhy et Marais, 2012).

Il existe toutefois quelques techniques et indices statistiques pouvant aider à diagnostiquer des problèmes locaux de respect des conditions d'utilisation du modèle de Rasch à multifacettes. Les indices d'ajustement des carrés moyens pondérés ou

[12] Il serait peut-être possible d'adapter les techniques de *multiway factor analysis* pour étudier ces matrices non rectangulaires, mais cela irait au-delà de nos visées, aucune étude utilisant le modèle de Rasch à multifacettes n'ayant utilisé de telles techniques (Kroonenberg, 2008).

non (*infit* ou *outfit*) peuvent aider à identifier des problèmes potentiels, mais ces indices ne ciblent pas précisément l'une ou l'autre des conditions d'utilisation et des valeurs aberrantes peuvent être causées par divers types de problèmes. La situation est compliquée, car beaucoup de travail a été accompli quant aux propriétés distributionnelles des indices utilisés avec des données dichotomiques (Béland, 2015 ; Karabatsos, 2000 ; Kreiner et Christensen, 2016 ; Raîche, Magis, Blais et Brochu, 2013 ; R. M. Smith, 1988, 1991 ; Wu et Adams, 2013). Par contre, peu d'études ont étudié les propriétés distributionnelles pour des données polychotomiques (Seol, 2016 ; A. B. Smith, Rush, Fallowfield, Velikova et Sharpe, 2008). De surcroît, aucune étude n'a examiné les propriétés distributionnelles des indices d'ajustement des éléments de la facette « examinateur » dans un modèle de Rasch à multifacettes. Ces indices de carrés moyens pondérés et non pondérés seront néanmoins utilisés et les valeurs référentielles de 0,5 à 1,5 pour les carrés moyens seront retenues pour identifier les examinateurs potentiellement problématiques, car cette thèse ne cherche pas à utiliser les évaluations des examinateurs pour mesurer précisément les performances des candidats dans le but de prendre une décision affectant ceux-ci, mais bien à examiner l'évolution temporelle du niveau de sévérité des examinateurs eux-mêmes. Il semble donc acceptable d'utiliser les bornes clémentes, moins restrictives, suggérées par Linacre (2002, 2017a).

Pour tenter de détecter des problèmes importants de dépendance locale entre les examinateurs, l'indice Rasch-kappa sera utilisé (Eckes, 2011 ; Linacre, 2017a). Cet indice repose sur la différence entre le nombre observé de cas où deux examinateurs ont attribué au même candidat la même note au même critère et le nombre attendu de fois où, étant donné les valeurs des paramètres impliqués du modèle (niveau d'habileté, niveau de difficulté des critères d'évaluation, niveau de sévérité des examinateurs et seuil de difficulté de chaque catégorie de l'échelle d'appréciation), deux examinateurs auraient dû accorder la même note. L'équation 7 illustre le calcul de l'indice :

$$Rasch\text{-}kappa = \frac{(\%\ Observé - \%\ Attendu)}{(100 - \%\ Attendu)} \quad (7)$$

Sa valeur attendue, si les examinateurs sont parfaitement indépendants et que leur degré d'accord correspond au degré attendu par le modèle, est de 0. Des valeurs faibles, proches de 0, sont ainsi un signe du respect de l'indépendance locale. Une valeur positive indique une dépendance entre les notes accordées par les examinateurs dont l'ampleur est liée à la valeur de l'indice. Sa valeur maximale théorique est de 1 et sa valeur minimale théorique est approximativement de -99, mais précisons que ces bornes représentent des cas extrêmes, peu susceptibles de se produire dans la pratique : respectivement un pourcentage observé de 100 pour le maximum et un pourcentage observé de 0 couplé à un pourcentage attendu de 99.

Alors que les deux types d'indice précédents concernent l'ajustement local des données au modèle, le pourcentage de résidus standardisés supérieurs à certaines valeurs absolues sera utilisé pour vérifier l'adéquation globale des données au modèle. Puisque la distribution des résidus standardisés approxime la distribution d'une loi normale, la littérature considère qu'au plus 5 % des résidus standardisés devrait avoir une valeur absolue égale ou supérieure à 2 et qu'au plus 1 % une valeur absolue égale ou supérieure à 3 (Eckes, 2005, 2011 ; Linacre, 2017a, R. M. Smith, 1988). Ces bornes seront donc utilisées pour juger de l'adéquation et de la qualité des mesures obtenues.

La condition d'unidimensionnalité sera étudiée, mais sous un angle particulier. Ce n'est pas l'unidimensionnalité psychométrique de la note totale qui sera ici vérifiée, cette dernière étant tenue pour acquise puisque, dans la situation réelle d'évaluation, une note totale est calculée pour chaque candidat et cette note est utilisée pour attribuer un niveau de compétence au candidat. Il s'agit plutôt de l'unidimensionnalité des notes accordées par les examinateurs afin de s'assurer que

certains des examinateurs ne définissent pas conjointement une dimension secondaire qui pourrait biaiser les notes totales. Cette démarche s'apparente au concept de fonctionnement différentiel en ce qu'il faut vérifier si les examinateurs tendent à utiliser les critères d'une manière similaire lorsqu'ils notent les performances des candidats. Il faut, pour ce faire, utiliser le résultat d'une analyse avec l'une des modélisations décrites ci-dessus et en arranger les résidus standardisés de manière à produire la matrice rectangulaire 39 996 × 20 suivante : chaque ligne correspond à une combinaison candidat (3 333), critère (12), note (1) accordée par un examinateur et chaque colonne à un examinateur (20). La facette factice « temps » est ignorée. Cette matrice de données sera analysée à l'aide d'un modèle de Rasch pour données polychotomiques à deux facettes (les candidats et les critères ayant été fusionnés). La dimension correspondant au modèle de Rasch ayant été « extraite », les résidus standardisés de cette analyse seront soumis à une analyse en composantes principales afin de vérifier si les examinateurs se regroupent en une dimension secondaire dont les mesures seraient faiblement corrélées aux mesures estimées selon le modèle de la « dimension Rasch » (Eckes, 2011 ; Linacre, 1998, 2017b). Les balises utilisées pour interpréter l'importance et le risque potentiel posés par les dimensions secondaires seront, en premier lieu, l'ampleur des valeurs propres de ces dimensions, qui devront être inférieures à 2 (Linacre, 1998, 2017b). En second lieu, les corrélations désatténuées[13] entre les mesures d'une dimension secondaire et les mesures de la dimension Rasch devraient être égales ou supérieures à 0,95 afin que cette dimension secondaire puisse être considérée comme redondante et que l'unidimensionnalité soit ainsi respectée (voir Linacre, 2017b pour les détails techniques).

L'utilisation des procédures décrites ci-dessus permettra de juger de la qualité des mesures estimées pour le niveau de sévérité des examinateurs et d'éventuels changements à apporter aux données (élimination de certaines évaluations faites par

[13] Traduction libre de *disattenuated*. Voir le glossaire.

certains examinateurs, voire un retrait complet d'un examinateur en cas d'indices d'ajustement catastrophiques) avant de pouvoir, à la deuxième étape d'analyse, utiliser les estimés des niveaux de sévérité comme variable dépendante de séries chronologiques.

3.4.2 Séries chronologiques des niveaux de sévérité des examinateurs

Une série chronologique est une série temporellement ordonnée de mesures d'une variable X_t où t représente chaque intervalle temporel successif. Une série peut être continue, par exemple dans le cas d'une onde sonore, auquel cas t est un nombre réel. Elle peut aussi être discrète, dans le cas d'un phénomène mesuré à des intervalles de temps définis, auquel cas t est un nombre naturel. Il existe des séries chronologiques univariées et multivariées, lorsqu'il y a au moins deux variables X_{tij} mesurées sur la même période de temps. Pour cette thèse, les séries chronologiques utilisées seront discrètes univariées, puisqu'il n'est pas théoriquement sensé de supposer qu'un même processus stochastique soit responsable de l'évolution du niveau de sévérité de plusieurs personnes, le niveau de sévérité étant le résultat d'un processus évaluatif psychologique propre à chaque personne.

Le but général de la modélisation par séries chronologiques est de décrire l'évolution temporelle de la variable X_t et, éventuellement, d'en prédire les valeurs pour les intervalles de temps à venir. Les séries chronologiques ont potentiellement 4 composantes : la tendance, qui représente la croissance ou la décroissance de X à long terme, la saisonnalité, qui représente les changements de X à moyen terme, sur une période se répétant régulièrement au sein des données collectées, le cycle, qui représente des fluctuations régulières, mais sur une période inconnue et les fluctuations aléatoires, qui représentent les changements irréguliers. Nonobstant les éléments techniques pouvant différer, l'idée de base d'une série chronologique est d'identifier un modèle mathématique décrivant le mieux possible les données

observées, c'est-à-dire le processus stochastique responsable de la génération des données collectées, puisque l'on suppose que les valeurs de X à un temps t dépendent en partie des valeurs de X aux temps antérieurs à t. La méthode générale de Box-Jenkins servira d'heuristique pour l'analyse des séries chronologiques (Box, Jenkins et Reinsel, 1994). Cette méthode suppose que les données peuvent être décrites par un modèle de la classe des modèles Autorégressifs à Moyennes Mobiles Intégrés (AMMI – ARIMA en anglais) et elle propose la démarche itérative suivante : identifier le modèle stochastique approprié, estimer les paramètres du modèle, vérifier l'adéquation du modèle aux données et, si le modèle semble optimal, utiliser le modèle retenu pour faire des prévisions, sinon revenir à l'identification et suivre de nouveau les étapes. Pour cette thèse, les modèles AMMI seront retenus plutôt que leur équivalent intégrant des paramètres saisonniers (AMMIS), car il n'y a aucune raison théorique de supposer la présence de saisonnalité dans les données et la présence d'une telle composante serait étonnante. Les modèles de la classe AMMI combinent un modèle autorégressif (p) et un modèle à moyenne mobile (q) pouvant être représentés par les équations respectives suivantes :

$$A\,x_t = c + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t \qquad (8)$$

où c est une constante, ε du bruit blanc de moyenne 0 et de variance finie, ϕ les coefficients de régression et p l'ordre de la fonction d'autorégression et

$$MM\,x_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \qquad (9)$$

où μ est la moyenne de la série chronologique, ε du bruit blanc de moyenne 0 et de variance finie, θ les coefficients de régression et q l'ordre de la fonction de moyenne mobile. En d'autres termes, le modèle autorégressif stipule que la valeur de X à un temps t est fonction des valeurs de X à un temps antérieur à t et allant jusqu'à $t - p$ et le modèle à moyenne mobile stipule que la valeur de X à un temps t est fonction des valeurs de bruit blanc antérieures à t allant jusqu'à $t - q$. Le modèle complet AMM est représenté par l'équation suivante, plus compacte :

$$AMM\ x_t = \varepsilon_t + c + \sum_{i=p}^{p} \phi_i x_{t-i} + \sum_{j=q}^{q} \theta_j \varepsilon_{t-j} \qquad (10)$$

où les paramètres sont identiques aux paramètres identifiés précédemment. Si la série originale a été différenciée, un 3e paramètre s'ajoute au modèle, le paramètre *d* indiquant l'ordre de la différenciation, soit le nombre de fois où la série originale a été différenciée et le modèle AMM devient un modèle AMMI. Son équation, impliquant l'opérateur de décalage[14] (*L*) et l'opérateur de différenciation (∇), ne sera pas représentée, puisqu'il existe une identité mathématique entre ce modèle et le modèle AMM, mais les détails mathématiques, les théorèmes et démonstrations pertinentes se trouvent dans Brockwell et Davis (2002), Girard (2011) et Thibodeau (2011). Le modèle est noté de la manière suivante : AMMI (*p*,*d*,*q*), où les lettres renvoient aux paramètres décrits ci-dessus.

3.4.2.1 Vérification des conditions d'utilisation de la méthode Box-Jenkins

Cette méthode a quatre conditions d'utilisation ou spécification (Brockwell et Davis, 2002 ; Pankratz, 1983). Premièrement, les relations modélisées doivent être linéaires. Le respect de cette condition sera testé de deux manières. Premièrement, en vérifiant les graphiques en nuage de points des données X_t et X_{t-l}, où l représente le décalage, et ce pour tout $2 \leq l \leq 5$. Le choix de 5 est justifié par le fait qu'il est très improbable que des relations de dépendance à long terme existent pour ces données, le niveau de sévérité d'un examinateur à un moment t n'étant probablement pas lié au niveau de sévérité à un moment $t - 6$ mois, ou $t - 60$ candidats évalués. Ensuite, plusieurs statistiques inférentielles seront utilisées, selon l'hypothèse nulle qu'elles testent. L'hypothèse nulle de linéarité sera directement testée par le test de blancheur (*white noise test*) de Teräsvirta, Lin et Granger (1993). Ensuite, puisque des variables indépendantes n'ont, par définition, aucune relation, linéaire ou non linéaire, il est

[14] Aussi appelé « opérateur délai » ou « opérateur retard » dans la littérature francophone.

possible de tester cette hypothèse pour montrer l'absence de relations non linéaires. L'hypothèse nulle selon laquelle les résidus du modèle optimal retenu sont des variables indépendantes identiquement distribuées sera testée par les tests de Ljung-Box (1978) et de BDS (Brock, Dechert et Sheinkman, 1987).

Deuxièmement, la série modélisée doit être faiblement stationnaire, c'est-à-dire qu'elle doit avoir une moyenne et une variance constantes, indépendantes du temps t et la covariance entre tout X_t et X_{t-l} doit aussi être indépendante de t. Une série faiblement stationnaire a une tendance et une saisonnalité nulles, mais elle peut avoir des cycles et des fluctuations irrégulières. L'hypothèse nulle selon laquelle le modèle retenu ou ses résidus sont faiblement stationnaires sera testée par les tests KPSS (Kwiatkowski, Phillips, Schmidt et Shin, 1992), augmenté de Dickey Fuller (Said et Dickey, 1984) et B de Bartlett pour bruit blanc (Bartlett, 1967). Pour les deux premières conditions d'utilisation, le nombre élevé de statistiques inférentielles retenues s'expliquent par le fait que ces statistiques ont des niveaux de puissance distincts selon les hypothèses alternatives possibles et que ces statistiques serviront à la fois à vérifier le respect de la condition de linéarité et de l'adéquation du modèle optimal retenu aux données. Si la série chronologique initiale n'est pas stationnaire, deux techniques pourront être utilisées pour stationnariser celle-ci : différencier la série ou la transformer (Brockwell et Davis, 2002). Différencier une série consiste en le fait de calculer la différence entre toutes les paires de données consécutives et d'utiliser ces différences comme données à modéliser. C'est la définition de l'opérateur de différenciation (∇), représenté par l'équation suivante :

$$\nabla x_t = x_t - x_{t-1} \tag{11}$$

Généralement, une ou deux différenciations suffisent à stationnariser une série chronologique. Si la série initiale est différenciée, il faut intégrer le modèle optimal retenu pour revenir à la série initiale, tel que mentionné à la section précédente. L'autre possibilité est de transformer les données, par exemple à l'aide d'une

transformation logarithmique ou racine carrée. Dans les deux cas, les statistiques inférentielles présentées au paragraphe précédent seront utilisées pour vérifier la stationnarité de la série différenciée ou transformée. Dans les deux cas, la valeur et la direction de la tendance seront estimées par l'ajustement d'un polynôme de degré n minimisant la somme des carrés des erreurs, ce qui permettra de décrire la tendance macroscopique du niveau de sévérité. Pour cette raison – décrire la tendance et non la modéliser – une tendance polynômiale sera préférée à une tendance par moyenne mobile, car une tendance polynômiale est plus à même de répondre à la question conceptuellement importante sous-jacente : « les examinateurs deviennent-ils plus ou moins sévères avec le temps? »

La troisième condition d'utilisation concerne le nombre d'observations minimales. Le nombre minimal dépend principalement de trois facteurs : la présence et la périodicité de la saisonnalité, le nombre de paramètres du modèle optimal ainsi que le rapport bruit/signal dans les données (Hyndman et Kostenko, 2007). Tel que mentionné ci-dessus, aucune composante saisonnière n'est prévue, ce qui diminue d'autant le nombre de paramètres potentiels du modèle optimal. Ainsi, il devrait être possible de modéliser de courtes séries et la borne inférieure sera fixée à 10 temps de mesure. La modélisation avec un aussi petit nombre de temps de mesure entraîne généralement des erreurs types très importantes, mais il se peut néanmoins que le modèle retenu apporte des informations importantes pour répondre aux questions de recherche. Dans tous les cas, il faudra s'assurer qu'il y aura au minimum 1 temps de mesure de plus qu'il n'y aura de paramètres estimés. Par conséquent, les niveaux de sévérité des 7 examinateurs ayant évalué moins de 100 candidats ne seront pas modélisés à l'aide de séries chronologiques.

La quatrième condition d'utilisation concerne les résidus. Ceux-ci doivent se comporter comme du bruit blanc, soit avoir une espérance de 0, une variance

constante et finie et une autocovariance nulle. Cette condition sera testée avec les statistiques inférentielles énumérées pour la deuxième condition d'utilisation ci-dessus.

3.4.2.2 Utilisation de la méthode Box-Jenkins

Pour identifier le modèle, les données seront représentées à l'aide d'un graphique chronologique. Cela permettra d'identifier d'éventuelles valeurs extrêmes et de décider s'il faut ou non transformer ou différencier la série chronologique originale. L'examen de ce graphique et des fonctions d'autocorrélation et d'autocorrélation partielle permettront d'identifier un premier modèle, qui sera mis à l'épreuve. La fonction d'autocorrélation calcule le coefficient de corrélation linéaire entre toutes les valeurs X_t et X_{t-l} et l'autocorrélation partielle calcule le coefficient de corrélation linéaire partielle entre toutes les valeurs X_t et X_{t-l} tout en contrôlant la covariance entre les valeurs intermédiaires situées entre X_t et X_{t-l} pour calculer la corrélation entre X_t et X_{t-l}. Les valeurs significativement différentes de 0, selon un intervalle de confiance à 95 %, aident à identifier des valeurs p et q plausibles, car différents patrons de valeurs sont associés à différents modèles AMM (Brockwell et Davis, 2002 ; Evans, 2003 ; Hyndman et Athanasopoulos, 2014). Pour identifier un modèle, il faut choisir les valeurs optimales des paramètres p, d et q pour ensuite estimer les valeurs des coefficients θ, ϕ ainsi que la variance du bruit blanc, σ^2. Le modèle ainsi estimé est comparé aux données observées et les différences entre les valeurs estimées et obtenues constituent les résidus. Ces résidus seront aussi analysés avec un graphique chronologique ainsi qu'avec les fonctions d'autocorrélation totale et partielle afin de s'assurer que ces résidus se comportent comme du bruit blanc, sans tendance et qu'ils ne sont pas corrélés. Si ce n'est pas le cas, une ou plusieurs modifications seront apportées à l'identification du modèle, jusqu'à ce qu'un modèle mène à des résidus se comportant comme du bruit blanc. Suivant les recommandations de Brockwell et Davis (2002) et de Hurvich et Tsai (1989) pour

choisir entre deux modèles concurrents dont les résidus se comportent comme du bruit blanc, le modèle minimisant la fonction du critère d'information d'Akaike corrigé (CIAc,) sera retenu comme modèle optimal. Ce modèle sera retenu comme la meilleure représentation de l'évolution temporelle du niveau de sévérité d'un examinateur, ce qui permettra d'atteindre les objectifs spécifiques de recherche énoncés à la fin du contexte théorique comme le montrent les sections suivantes.

3.4.2.3 Premier objectif spécifique de recherche

Pour rappel, cet objectif est formulé ainsi : « Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du nombre de candidats évalués et du temps chronologique ». Pour ce faire, les valeurs estimées des paramètres des modèles AMMI de séries chronologiques retenues seront présentées et commentées ainsi que les graphiques chronologiques de ces séries. Au total, 84 séries chronologiques univariées seront modélisées, soit 36 pour les 12 examinateurs ayant évalué au moins 100 candidats et 48 pour le cas où le temps est fonction du temps chronologique (voir les informations du tableau 3.2 pour les détails). En sus de l'information visuelle apportée par les graphiques chronologiques, la présence et l'importance d'une tendance, de cycles, des fluctuations irrégulières ainsi que d'éventuelles hétéroscédasticités seront les éléments importants permettant d'atteindre cet objectif. Les statistiques descriptives des séries chronologiques seront également présentées et analysées, de même que les corrélations croisées intraindividuelles entre les niveaux de sévérité A et B, A et L ainsi que B et L. La corrélation linéaire entre deux séries chronologiques au même temps t est la corrélation linéaire normale entre deux variables. La corrélation croisée est, elle, utilisée afin de savoir si une série chronologique Xt « mène » l'autre et si ses valeurs à un temps t sont liées aux valeurs de Yt, mais à un temps antérieur l. La corrélation croisée calcule le coefficient de corrélation linéaire entre les séries Xt et $Yt\text{-}l$ (Brockwell et Davis, 2002).

3.4.2.4 Deuxième objectif spécifique de recherche

Le deuxième objectif est formulé ainsi : « Comparer l'évolution du niveau de sévérité des examinateurs débutants et expérimentés ». Cet objectif sera atteint différemment selon la modélisation temporelle retenue. Dans le cas où le temps est découpé en fonction du nombre de candidats évalués, les niveaux de sévérité estimés provenant des analyses où chaque intervalle de temps correspond à 5 candidats évalués seront utilisés. L'évolution du niveau de sévérité des premières périodes ($t \leq 20$) sera analysée en détail et une attention particulière sera accordée à la présence éventuelle de maxima et de minima, de ruptures, de tendances locales différentes de la tendance globale et d'autres irrégularités potentielles dans les séries chronologiques des examinateurs débutants et expérimentés. Dans les 5 autres modélisations temporelles, où le temps est découpé en fonction du temps chronologique, l'évolution temporelle du niveau de sévérité des examinateurs débutants sera comparée à celle des examinateurs expérimentés présents. Plus précisément, l'évolution représentant l'évaluation de 100 candidats par le ou les examinateurs débutants sera étudiée – ce qui représentera un nombre variable de temps de mesure en fonction du nombre de candidats évalués par le ou les examinateurs débutants. Les éléments suivants seront étudiés : les écarts de niveaux de sévérité à chaque temps de mesure, la convergence ou à la divergence entre ces niveaux, la présence de valeur minimale ou maximale de niveau de sévérité pour cette période, la variance des niveaux de sévérité ainsi que la présence de tendances locales.

3.4.2.5 Troisième objectif spécifique de recherche

Le troisième objectif est ainsi formulé : « Comparer l'évolution du niveau de sévérité d'examinateurs travaillant ensemble ». Pour atteindre cet objectif, les séries chronologiques issues des analyses où le temps est découpé selon le temps chronologique, décrites à la section 3.4.1.2, seront utilisées. Pour chacune des 5 modélisations décrites à cette section, les corrélations croisées interindividuelles

seront calculées pour chaque paire de séries chronologiques, au décalage $l \pm 1$. Cela permettra de vérifier s'il y a ou non des liens entre l'évolution du niveau de sévérité de chaque paire d'examinateurs travaillant conjointement durant la même période de temps. Les représentations graphiques associées seront également examinées afin de s'assurer de l'absence de relations non linéaires.

Pour terminer, une analyse supplémentaire relevant à la fois du premier et du troisième objectif spécifique de recherche sera effectuée. Pour chacune des 6 modélisations temporelles de cette thèse, le rapport entre l'étendue intraindividuelle et interindividuelle des niveaux de sévérité des examinateurs sera calculé comme suit. L'étendue intraindividuelle, pour un examinateur d'une période donnée, est égale à la différence entre la valeur maximale de son niveau de sévérité et la valeur minimale. L'étendue interindividuelle est obtenue de la manière suivante : pour chaque temps t d'une période de temps, l'écart entre le niveau de sévérité maximal et minimal est calculé. L'étendue interindividuelle correspond au plus grand écart ainsi obtenu. Le rapport entre l'étendue intraindividuelle et interindividuelle est ensuite calculé à l'aide de la fraction suivante :

$$Rapport\ étendues = \frac{Étendue\ intraindividuelle}{Étendue\ interindividuelle} \tag{12}$$

Ce rapport permettra de voir, pour chaque examinateur, jusqu'à quel point son niveau de sévérité est temporellement stable lorsque comparé aux écarts observés entre les niveaux de sévérité des examinateurs.

3.4.3 Logiciels utilisés

Les analyses pour estimer les niveaux de sévérité des examinateurs seront effectuées avec le logiciel *Facets* (version 3.71.4, Linacre, 2014), sauf pour l'analyse de la dimensionnalité des examinateurs, qui sera faite avec *Winsteps* (version 3.90.2, Linacre, 2015). L'estimation des paramètres du modèle de Rasch à multifacettes sera

faite avec un estimateur par maximum de vraisemblance conjoint (Linacre, 1994). Toutes les analyses concernant le traitement des données et les séries chronologiques seront faites avec le logiciel R (version 3.3.1, R Core Team, 2013) et ses bibliothèques suivantes : *fNonlinear* (version 3010.78, Wuertz et Chalabi, 2015), *forecast* (version 7.2, Hyndman, 2016), *hwwntest* (version 1.3, Savchev et Nason, 2015), *locits* (version 1.7.1, Nason, 2016), *normwhn.test* (version 1.0, Wickham, 2015) et *tseries* (version 0.10-35, Trapletti, Hornik et LeBaron, 2016). L'estimation des paramètres du modèle AMMI sera faite avec un estimateur par maximum de vraisemblance.

3.5 Considérations éthiques

Puisque les données utilisées sont des données secondaires et que la quantité de données ne permet pas d'identifier un participant par recoupement, le *Comité d'éthique de la recherche avec ces êtres humains* de l'UQAM a jugé que l'approbation éthique n'était pas nécessaire (Covanti, courriel, 9 septembre 2016). Toutes les précautions en vigueur ont néanmoins été suivies. Les données sont anonymisées et confidentielles et seul l'auteur y a accès. Aucune information susceptible de permettre l'identification d'un participant précis n'est fournie et les données sociodémographiques sont réduites au strict minimum, si bien que les conséquences négatives potentielles pour les participants de l'utilisation de ces données à des fins de recherche sont quasi nulles. En revanche, la diffusion des résultats de cette thèse pourrait, éventuellement, mener à l'amélioration des pratiques de sélection et de formation des examinateurs d'examens à forts enjeux. En plus de faire l'objet de communications orales et écrites, les résultats pertinents seront transmis au centre de test, ce qui fournira de précieuses informations à ce dernier pour le contrôle de la qualité du travail de ses examinateurs, ce qui ne peut qu'améliorer le processus de validation continue des tests concernés.

CHAPITRE IV

RÉSULTATS

L'atteinte des objectifs spécifiques de cette thèse suppose deux étapes analytiques, dont la première n'est qu'une étape préalable. D'abord, l'étape préliminaire de l'utilisation du modèle de Rasch à multifacettes pour estimer les niveaux de sévérité des examinateurs à différents temps de mesure, puis, pour la seconde étape, l'utilisation de statistiques descriptives, de graphiques chronologiques et de la modélisation AMMI pour l'atteinte proprement dite des 3 objectifs spécifiques de recherche. Les résultats de la première étape seront présentés en premier, mais de manière succincte, puisque cette étape ne sert qu'à obtenir les valeurs estimées des niveaux de sévérité, valeurs utilisées à la seconde étape. Seuls les résultats pertinents à la vérification des conditions d'utilisation du modèle de Rasch à multifacettes, telles que définies à la section 3.4.1.3, seront présentés. Les niveaux de sévérité étant estimés pour 3 notes, A, B et L, les analyses de la première étape seront séparées en 3 ensembles distincts et les résultats seront également présentés séparément pour les 3 notes, et ce pour s'assurer que les 3 ensembles de données respectent les conditions d'utilisation du modèle de Rasch à multifacettes.

Rappelons que ces 3 notes renvoient à la division des 12 critères d'évaluation utilisés par les examinateurs. Les 3 premiers critères (A) servent à évaluer la performance d'un candidat lors de la première tâche communicative, les 3 critères suivants (B) évaluent la performance d'un candidat lors de la seconde tâche communicative et les 6 derniers critères (L) évaluent les qualités linguistiques du candidat. Ainsi, pour

chaque candidat, 3 notes sont générées et, conséquemment, le niveau de sévérité des examinateurs est étudié séparément pour chacune des 3 notes

Ensuite, les résultats de la seconde étape, soit l'atteinte des 3 objectifs spécifiques de recherche, seront présentés. Les 3 objectifs sont : 1) « Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du temps chronologique et du nombre de candidats évalués », 2) « Comparer l'évolution du niveau de sévérité des examinateurs débutants et expérimentés » et 3) « Comparer l'évolution du niveau de sévérité d'examinateurs travaillant ensemble ». Afin de faciliter la lecture des résultats, ceux-ci seront présentés par ensemble de données, tous les résultats pertinents à l'atteinte des 3 objectifs spécifiques étant détaillés pour chaque ensemble de données. Il y aura donc 6 sections à cette deuxième étape, puisque les niveaux de sévérité des examinateurs seront étudiés en fonction du nombre de candidats évalués (1 ensemble de données) et en fonction du temps chronologique (5 découpages différents). Notons que, pour l'ensemble de données provenant de l'étude du niveau de sévérité des examinateurs en fonction du nombre de candidats évalués, aucun résultat n'est présenté concernant le troisième objectif spécifique de recherche, car, pour ces données, chaque examinateur est étudié isolément. Pour les 5 ensembles de données où le temps est découpé de manière chronologique, les résultats concernant ce troisième objectif spécifique seront présentés dans la section intitulée : « Corrélations croisées interindividuelles ». De même, pour certains ensembles de données obtenus en fonction du temps chronologique, aucun résultat n'est présenté pour le deuxième objectif spécifique, car ces ensembles de données ne comportent aucun examinateur débutant.

4.1 Respect des conditions d'utilisation du modèle de Rasch à multifacettes

4.1.1 Respect des conditions d'utilisation pour les données A

Les données ici étudiées sont les données complètes, sans égard au temps. Il y a donc les 3 333 candidats, les 20 examinateurs et les 3 critères d'évaluation en une seule analyse et le tableau 4.1 montre les principales statistiques descriptives de la distribution des mesures, en logit, des paramètres d'habileté des candidats, de sévérité des examinateurs et de difficulté des critères d'évaluation. Le tableau 4.2 contient la distribution par quartiles des erreurs types des valeurs estimées des paramètres d'habileté des candidats et de sévérité des examinateurs.

Tableau 4.1
Statistiques descriptives des paramètres des candidats, des examinateurs et des critères d'évaluation, en logit, pour les données A

	min	max	étendue	*m*	*s*
Candidats	-12,68	7,15	19,83	0,00	2,64
Examinateurs	-2,24	-1,26	0,98	-1,73	0,24
Critères	-0,17	0,18	0,35	0,00	0,18

Tableau 4.2
Distribution par quartiles des erreurs types des valeurs estimées des paramètres des candidats et des examinateurs, pour les données A*

	min	25 %	md	75 %	max
Candidats	0,26	0,28	0,30	0,32	1,87
Examinateurs	0,01	0,02	0,03	0,06	0,10

* La facette « Critères d'évaluation » n'est pas présente, car les 3 critères ont tous la même erreur type, soit 0,01

Rappelons que les facettes « Candidats » et « Critères » ont une moyenne de 0 par définition, puisque cette contrainte sert à ancrer les estimations à un point 0 unique pour les 3 facettes. Constatons que les 3 333 candidats on un niveau d'habileté ayant une grande étendue, un écart type élevé et que 94,8 % des candidats ont un niveau d'habileté situé à ±2 écarts types de la moyenne. Cette grande étendue n'est pas

surprenante, puisqu'à partir de 2012, presque tous les candidats à l'immigration devaient faire ce test, peu importe leur niveau d'habileté. Il y avait donc des candidats sans aucune connaissance du français et d'autres qui étaient francophones de naissance ou de scolarité avec une éloquence très développée. Cela représente un continuum de niveaux d'habileté très important expliquant l'étendue de près de 20 logits des mesures des candidats. Mentionnons aussi que les valeurs de ces tableaux incluent les notes brutes extrêmes, c'est-à-dire les candidats ayant reçu la note minimale ou maximale pour les 3 critères d'évaluation A. Cela fait que le niveau d'habileté de ces candidats est inestimable. Les valeurs minimales et maximales présentées sont des projections obtenues à l'aide d'une approximation détaillée dans Wright (1998). Les valeurs minimales et maximales non extrêmes sont respectivement -11,20 et 5,87 logits, pour une étendue de 17,07 logits. Les erreurs types des valeurs estimées d'habileté sont approximativement distribuées selon une parabole qui aurait son minimum à la valeur moyenne du paramètre d'habileté. En d'autres termes, plus une valeur estimée d'habileté est proche du centre de cette distribution, plus l'estimation est précise et l'erreur type petite. Inversement, les valeurs estimées d'habileté proches du minimum (-12,70) ou du maximum (7,10) sont très imprécises et les erreurs types sont élevées. Les erreurs types des valeurs extrêmes sont, elles, très grandes (1,84 à 1,87), mais elles sont aussi des projections obtenues par approximation (Wright, 1998)[15]. L'erreur type maximale des valeurs estimées non extrêmes est de 1,09, une valeur plus raisonnable. Cela dit, 75 % des erreurs types sont égales ou inférieures à 0,32, ce qui fait, étant donné l'étendue de la distribution des niveaux d'habileté, que ces estimations sont assez précises.

Les examinateurs, eux, ont un niveau de sévérité global homogène, avec une étendue d'environ 1 logit et un écart type assez faible. Les erreurs types des valeurs estimées du niveau de sévérité sont inversement proportionnelles au nombre de candidats

[15] Ce qui est l'option par défaut du logiciel *Facets*.

évalués par les examinateurs ; plus un examinateur a évalué de candidats, plus l'erreur type de la valeur estimée de son niveau de sévérité est petite. La moitié des examinateurs, ceux ayant le plus travaillé, a une erreur type très petite (de 0,01 à 0,03) et même l'examinateur ayant le moins travaillé (9H) a une erreur type assez faible (0,10). Les 3 critères d'évaluation ont aussi un niveau de difficulté homogène et ils ont tous la même erreur type, soit 0,01 logit, ce qui est normal puisque plus de 6 000 notes ont été accordées. Suite à l'étude de ces statistiques descriptives, les analyses annoncées à la section « Méthodologie » ont donc été faites pour vérifier le respect des conditions d'utilisation du modèle de Rasch à multifacettes pour cet ensemble de données.

Les données A sont globalement bien ajustées au modèle, avec un pourcentage de résidus standardisés supérieurs à 3 ou inférieurs à -3 égal à 0,67 % et un pourcentage supérieurs à 2 ou inférieurs à -2 égal à 4,55 %, soit des pourcentages à l'intérieur des valeurs maximales recommandées de 1% et 5 %. Le tableau 4.3 montre les valeurs des indices d'ajustement des examinateurs, soit les carrés moyens pondérés (*infit*) ou non pondérés (*outfit*) ainsi que l'indice Rasch-kappa. Le niveau de sévérité global, en logit, est également donné, à titre d'information supplémentaire.

Tableau 4.3
Indices d'ajustement et niveau de sévérité des examinateurs pour les données A

Examinateur	Niveau de sévérité	Carré moyen non pondéré	Carré moyen pondéré	Indice Rasch-kappa
1H	-1,41	1,13	1,12	-0,05
2F	-2,03	0,71	0,71	-0,03
3F	-1,71	1,24	1,23	-0,05
4F	-1,86	1,14	1,14	-0,06
5F	-1,51	0,96	1,00	-0,04
6F	-1,56	1,11	1,14	-0,04
7F	-1,50	1,23	1,24	-0,06
8F	-2,24	1,32	1,35	-0,06
9H	-1,26	1,34	1,39	**-0,18**
10F	-1,87	**1,97**	**1,80**	-0,03
11H	-1,79	1,34	1,45	**-0,13**
12F	-1,81	1,12	1,13	-0,01
13H	-1,46	0,77	0,78	-0,02
14F	-1,81	1,41	**1,63**	-0,05
15F	-1,94	1,07	1,16	0,00
16H	-2,12	**1,54**	**1,82**	-0,07
17F	-1,75	**0,47**	**0,49**	0,03
18F	-1,78	1,11	1,14	-0,07
19H	-1,64	0,54	0,54	**0,10**
20F	-1,62	**0,48**	**0,49**	0,04

Les valeurs en **gras** sont à l'extérieur des bornes recommandées

Il y a 5 examinateurs ayant des indices d'ajustement à l'extérieur des bornes retenues pour cette thèse, soit de 0,5 à 1,5 (10F, 14F, 16H, 17F et 20F), mais ce n'est pas vraiment un problème. Les examinateurs 10F, 14F et 16H n'ont évalué, respectivement, que 27, 30 et 64 candidats, ce qui est négligeable, puisque cela ne représente que 1,8 % de toutes les évaluations ayant été faites. Leur mauvais ajustement relatif au modèle ne pose donc qu'un très faible risque d'altérer les résultats subséquents. Les examinatrices[16] 17F et 20F ont aussi des indices à l'extérieur des bornes, mais c'est un artefact statistique causé par les indices élevés des 3 examinateurs susmentionnés ; comme les carrés moyens sont contraints d'avoir

[16] Les lettres « H/F » renvoient au genre de l'examinateur. Le mot est par conséquent au féminin.

une moyenne de 1,00, la présence d'indices élevés (10F et 16H) est directement responsable des valeurs très faibles, tout juste sous les bornes de 17F et 20F. Ces examinatrices peuvent donc être considérées comme bien ajustées au modèle. Il ne semble pas y avoir de violation importante de l'indépendance locale, 3 examinateurs ayant un indice Rasch-kappa un peu élevé (19H), ce qui montre une légère dépendance, ou faible (9H et 11H), ce qui montre, au contraire, une indépendance supérieure à celle attendue par le modèle. Ces 3 examinateurs ont tous évalué moins de 100 candidats, ce qui fait que cette très faible violation des conditions d'utilisation n'a vraisemblablement pas d'impact sur les analyses et résultats ultérieurs.

Finalement, le tableau 4.4 présente les résultats de l'analyse en composantes principales des résidus standardisés. Rappelons, pour en faciliter la compréhension, que cette analyse traite les combinaisons candidat + critère d'évaluation comme les sujets d'un ensemble de données traditionnel et chaque examinateur comme un item. Les examinateurs (les « items ») sont standardisés de manière à ce que chacun ait une valeur propre de 1. La valeur propre totale de la variance inexpliquée est donc nécessairement égale au nombre d'examinateurs dans l'analyse, ici 20. Le but de l'analyse est de voir si au moins deux « items » (examinateurs) forment une dimension secondaire d'une valeur propre substantielle, c'est-à-dire au minimum égale à 2, ce qui indiquerait que ces examinateurs « formant une dimension secondaire » tendraient à accorder des notes différentes par rapport aux autres examinateurs. Cela revient à vérifier si la facette des examinateurs respecte la condition d'utilisation d'unidimensionnalité.

Tableau 4.4
Valeurs propres de l'analyse en composantes principales des données A

	Valeur propre	% observé	% attendu
Variance totale	323,8	100	100
Variance expliquée par la 1ère dimension : Rasch	303,8	93,8	93,3
Variance inexpliquée totale	20,0	6,2	6,7
Variance 2e dimension	1,6	0,5	-
Variance 3e dimension	1,4	0,4	-
Variance 4e dimension	1,4	0,4	-
Variance 5e dimension	1,2	0,4	-

Pour les données A, les résultats de l'analyse montrent que la deuxième dimension n'explique que 0,49 % de la variance totale (1,6 sur 323,8) et sa valeur propre est inférieure à 2, ce qui signifie que cette dimension ne compte pas même pour 2 examinateurs, puisque l'analyse en composantes principales est faite sur les résidus standardisés où chaque examinateur vaut 1 unité de valeur propre. De surcroît, si l'on considère deux groupes d'examinateurs ayant les saturations les plus élevées, donc les examinateurs les plus susceptibles d'accorder des notes d'une manière différente, les corrélations désatténuées entre les niveaux d'habileté estimés par ces deux groupes sont égales ou supérieures à 0,95, ce qui tend à montrer que les 20 examinateurs ont une compréhension commune des critères d'évaluation et qu'ils respectent la condition d'unidimensionnalité. Les résultats de cette analyse permettent donc d'affirmer que la condition d'utilisation d'unidimensionnalité est respectée. Les résultats des analyses susmentionnées faites pour vérifier les conditions d'utilisation du modèle de Rasch à multifacettes montrent qu'il n'y a pas de violation importante de ces conditions d'utilisation pour les données A. Les valeurs estimées du niveau de sévérité des examinateurs pour ces données peuvent par conséquent être utilisées pour atteindre les 3 objectifs spécifiques de cette thèse.

4.1.2 Respect des conditions d'utilisation pour les données B

Le tableau 4.5 montre les principales statistiques descriptives de la distribution des mesures, en logit, des paramètres d'habileté des candidats, de sévérité des examinateurs et de difficulté des critères d'évaluation. Le tableau 4.6 présente la distribution en quartiles des erreurs types des valeurs estimées des paramètres d'habileté et de sévérité.

Tableau 4.5
Statistiques descriptives des paramètres des candidats, des examinateurs et des critères d'évaluation, en logit, pour les données B

	min	max	étendue	*m*	*s*
Candidats	-9,78	6,84	16,62	0,00	2,41
Examinateurs	-1,13	-0,68	0,45	-0,95	0,13
Critères	-0,33	0,37	0,70	0,00	0,35

Tableau 4.6
Distribution par quartiles des erreurs types des valeurs estimées des paramètres des candidats et des examinateurs, pour les données B*

	min	25 %	md	75 %	max
Candidats	0,21	0,24	0,27	0,31	1,86
Examinateurs	0,01	0,02	0,03	0,05	0,09

* La facette « Critères d'évaluation » n'est pas présente, car les 3 critères ont tous la même erreur type, soit 0,01

Les 3 333 candidats ont un niveau d'habileté ayant une grande étendue, bien qu'un peu plus petite que l'étendue pour les données A, un écart type élevé et 94,8 % des candidats ont un niveau d'habileté situé à ±2 écarts types de la moyenne, un pourcentage identique au pourcentage des données A. Les valeurs du tableau incluent les valeurs extrêmes approximées. Les valeurs estimées non extrêmes minimale et maximale sont de -8,47 et 5,60 logits, pour une étendue de 14,07 logits. Les erreurs types de ces valeurs sont également très proches des erreurs types des données A, avec essentiellement les mêmes valeurs. L'erreur type maximale pour les valeurs

estimées non extrêmes est de 1,07. Les examinateurs, eux, ont un niveau de sévérité global très homogène, avec une étendue minuscule d'à peine 0,45 logit et un écart type également minuscule. Ces estimations sont précises, les erreurs types des valeurs estimées de niveau de sévérité étant égales ou inférieures à 0,05 pour 15 examinateurs sur 20. Les 3 critères d'évaluation ont aussi un niveau de difficulté homogène et une étendue plus importante que les examinateurs ; l'erreur type des valeurs estimées des 3 critères d'évaluation est égale à 0,01. Les analyses ont donc été faites pour vérifier le respect des conditions d'utilisation du modèle de Rasch à multifacettes pour cet ensemble de données.

Les données B sont globalement bien ajustées au modèle, avec un pourcentage de résidus standardisés supérieurs à 3 ou inférieurs à -3 égal à 0,77 % et un pourcentage supérieurs à 2 ou inférieurs à -2 égal à 4,71 %. Le tableau 4.7 montre les valeurs des indices d'ajustement des examinateurs ainsi que leur niveau de sévérité.

Tableau 4.7
Indices d'ajustement et niveau de sévérité des examinateurs pour les données B

Examinateur	Niveau de sévérité	Indice ajustement non pondéré	Indice ajustement pondéré	Indice Rasch-kappa
1H	-0,83	1,20	1,16	-0,04
2F	-1,12	0,89	0,85	**-0,12**
3F	-1,06	1,49	1,47	-0,07
4F	-1,10	1,06	1,06	-0,04
5F	-0,88	1,04	1,09	-0,05
6F	-0,85	**1,67**	**1,78**	-0,02
7F	-0,87	1,09	1,09	-0,06
8F	-1,01	1,22	1,20	-0,05
9H	-0,68	**1,67**	**1,94**	**-0,16**
10F	-1,01	1,37	1,25	0,00
11H	-0,78	1,19	1,28	-0,05
12F	-0,84	0,92	0,90	0,02
13H	-0,93	0,77	0,78	0,00
14F	-0,84	1,35	1,42	**0,12**
15F	-1,05	1,12	1,21	-0,01
16H	-0,88	**1,84**	**2,35**	0,00
17F	-1,04	**0,46**	**0,49**	0,00
18F	-1,13	1,04	1,17	0,03
19H	-1,09	0,68	0,65	**0,10**
20F	-0,95	**0,48**	0,50	0,05

Les valeurs en **gras** sont à l'extérieur des bornes recommandées

Il y a 5 examinateurs ayant des carrés moyens à l'extérieur des bornes. Encore une fois, ces problèmes sont mineurs, puisque 2 de ces examinateurs ont évalué peu de candidats, 19 (9H) et 64 (16H) et que les carrés moyens inférieurs à 0,5 des examinatrices 17F et 20F sont causés par les valeurs très élevées de 16H et 9H. Reste le cas de 6F, qui a tout de même évalué 152 candidats, mais de manière très diffuse, sur une période de 26 mois. Comme les valeurs de ses carrés moyens ne sont pas trop éloignées de 1,5, les données de cette examinatrice ont été retenues pour les analyses subséquentes. Les valeurs de l'indice Rasch-kappa sont acceptables, avec seulement 4 examinateurs ayant des valeurs inférieures à -0,1 ou supérieures à 0,1, mais ces 4 examinateurs (2F, 9H, 14F et 19H) ont tous évalué moins de 100 candidats. Il n'y a

donc pas de violation de l'indépendance locale. Pour terminer, le tableau 4.8 présente les résultats de l'analyse en composantes principales des résidus standardisés.

Tableau 4.8
Valeurs propres de l'analyse en composantes principales des données B

	Valeur propre	% observé	% attendu
Variance totale	360,5	100	100
Variance expliquée par la 1ère dimension : Rasch	340,5	94,5	93,8
Variance inexpliquée totale	20,0	5,5	6,2
Variance 2e dimension	1,6	0,4	-
Variance 3e dimension	1,4	0,4	-
Variance 4e dimension	1,3	0,4	-
Variance 5e dimension	1,2	0,3	-

Ces résultats sont quasiment identiques aux résultats de l'analyse pour les données A. Les corrélations désatténuées entre les valeurs des niveaux d'habileté estimées par différents groupes d'examinateurs sont aussi, pour les données B, égales ou supérieures à 0,95. Ces résultats mènent à la même conclusion : il n'y a pas de violation notoire de l'unidimensionnalité. Il est donc possible d'affirmer, étant donné les résultats susmentionnés, que les données B respectent de manière satisfaisante les conditions d'utilisation du modèle de Rasch à multifacettes.

4.1.3 Respect des conditions d'utilisation pour les données L

Le tableau 4.9 montre les principales statistiques descriptives de la distribution des mesures, en logit, des paramètres d'habileté des candidats, de sévérité des examinateurs et de difficulté des critères d'évaluation. Le tableau 4.10, lui, présente la distribution en quartiles des erreurs types des valeurs estimées des paramètres d'habileté et de sévérité.

Tableau 4.9
Statistiques descriptives des paramètres des candidats, des examinateurs et des critères d'évaluation, en logit, pour les données L

	min	max	étendue	*m*	*s*
Candidats	-13,07	7,52	20,59	1,31	4,00
Examinateurs	-2,13	-1,35	0,78	-1,90	0,18
Critères	-0,30	0,26	0,56	0,00	0,23

Tableau 4.10
Distribution par quartiles des erreurs types des valeurs estimées des paramètres des candidats et des examinateurs, pour les données L*

	min	25 %	md	75 %	max
Candidats	0,19	0,21	0,22	0,44	1,85
Examinateurs	0,01	0,02	0,03	0,04	0,08

* La facette « Critères d'évaluation » n'est pas présente, car les 6 critères ont tous la même erreur type, soit 0,01

Notons premièrement que, contrairement aux données A et B, la facette « Candidats » n'a pas une moyenne de 0, mais bien une moyenne de 1,31. Ceci est dû à la présence de nombreux candidats (655) ayant une note extrême, minimale ou maximale, pour les 6 critères d'évaluation de la langue. Les 2 678 autres candidats ont bien une moyenne de 0, mais les valeurs estimées des 655 niveaux d'habileté « extrêmes » viennent gonfler la moyenne des 3 333 candidats, car il y a beaucoup plus de candidats ayant une note maximale (640) qu'une note minimale (15). Conséquemment, les 3 333 candidats ont un niveau d'habileté ayant une grande étendue, aussi élevée que l'étendue pour les données A, un écart type élevé et 97,7 % des candidats ont un niveau d'habileté situé à ±2 écarts types de la moyenne, un pourcentage très élevé. Le tableau 4.9 inclut les valeurs approximées des niveaux d'habileté des candidats ayant eu une note minimale ou maximale, ce qui augmente l'étendue. Les valeurs estimées minimales et maximales du niveau d'habileté des candidats ayant eu une note non extrême sont de -10,85 et 6,03 logits, pour une étendue de 16,88 logits. Les erreurs types de ces valeurs estimées sont plus petites

pour les données L que pour les données A et B, ce qui est normal puisque les données L comptent 6 items et non 3 comme les données A et B, mais seulement pour la moitié des erreurs types. Si les valeurs des deux premiers quartiles sont inférieures aux valeurs des quartiles correspondant pour les données A et B et le maximum similaire, le troisième quartile des données L est supérieur d'environ 0,10 aux troisièmes quartiles des données A et B, ceci dû au très grand nombre de candidats ayant une note maximale et, par conséquent, une erreur type très élevée. Ceci explique les différences dans la distribution par quartiles des erreurs types des données L par rapport à ces distributions pour les données A et B. L'erreur type la plus élevée pour un niveau d'habileté estimé à partir d'une note non extrême est de 1,01.

Les examinateurs, eux, ont un niveau de sévérité global très homogène, avec une étendue minuscule d'à peine 0,45 logit et un écart type également minuscule. Logiquement, les erreurs types de ces valeurs estimées sont très petites, 75 % des examinateurs ayant une erreur type égale ou inférieure à 0,04. Les 6 critères d'évaluation ont aussi un niveau de difficulté homogène et une étendue plus importante que les examinateurs ; encore une fois, l'erreur type des 6 critères d'évaluation est égale à 0,01. Les analyses ont donc été faites pour vérifier le respect des conditions d'utilisation du modèle de Rasch à multifacettes pour cet ensemble de données.

Les données L sont bien ajustées au modèle, avec un pourcentage de résidus standardisés supérieurs à 3 ou inférieurs à -3 égal à 0,71 % et un pourcentage supérieurs à 2 ou inférieurs à -2 égal à 3,64 %. Le tableau 4.11 montre les valeurs des indices d'ajustement des examinateurs.

Tableau 4.11
Indices d'ajustement et niveau de sévérité des examinateurs pour les données L

Examinateur	Niveau de sévérité	Indice ajustement non pondéré	Indice ajustement pondéré	Indice Rasch-kappa
1H	-1,93	1,31	1,28	-0,03
2F	-1,87	0,87	0,85	0,00
3F	-2,01	**1,66**	**1,59**	-0,07
4F	-2,08	0,98	1,03	-0,02
5F	-1,97	0,92	0,93	0,00
6F	-1,73	1,14	1,17	0,03
7F	-1,86	1,09	1,14	-0,04
8F	-2,13	1,03	1,07	-0,09
9H	-1,35	1,40	**1,59**	-0,02
10F	-1,99	1,16	1,26	**-0,13**
11H	-1,62	1,20	1,24	-0,05
12F	-1,83	1,01	1,03	0,07
13H	-2,06	0,93	0,95	0,03
14F	-2,08	0,77	0,79	0,01
15F	-1,96	1,25	1,30	0,01
16H	-1,92	**1,68**	**1,84**	**-0,10**
17F	-1,98	0,69	0,68	0,05
18F	-1,98	1,10	1,17	0,04
19H	-1,80	0,65	0,70	**0,10**
20F	-1,91	0,56	0,59	0,08

Les valeurs en **gras** sont à l'extérieur des bornes recommandées

Toutes les valeurs à l'extérieur des bornes recommandées, que ce soit pour les carrés moyens ou l'indice Rasch-kappa, appartiennent à des examinateurs ayant évalué très peu de candidats, soit 45 (3F), 19 (9H), 27 (10F), 64 (16H) et 89 (19H). Par conséquent, ces indices légèrement au-delà des bornes recommandées ne posent pas problème pour la qualité des données L lors des analyses subséquentes. Les résultats de l'analyse en composantes principales des résidus standardisés sont présentés dans le tableau 4.12.

Tableau 4.12
Valeurs propres de l'analyse en composantes principales des données L

	Valeur propre	% observé	% attendu
Variance totale	554,4	100	100
Variance expliquée par la 1ère dimension : Rasch	534,4	96,4	95,7
Variance inexpliquée totale	20,0	3,6	4,3
Variance 2e dimension	1,6	0,3	-
Variance 3e dimension	1,5	0,3	-
Variance 4e dimension	1,4	0,3	-
Variance 5e dimension	1,3	0,2	-

Ces résultats sont similaires aux résultats de ces analyses pour les données A et B et confirment que la condition d'utilisation d'unidimensionnalité est respectée, d'autant plus que les corrélations désatténuées entre les valeurs des niveaux d'habileté estimées par différents groupes d'examinateurs sont, pour les données L, égales ou supérieures à 0,98. Considérant l'ensemble des résultats présentés ci-dessus, les données L respectent de manière satisfaisante les conditions d'utilisation du modèle Rasch à multifacettes.

4.1.4 Conclusion de la 1re étape analytique

Les trois ensembles de données respectent largement les conditions d'utilisation du modèle de Rasch à multifacettes et les quelques cas de mauvais ajustement des données au modèle sont limités et concernent généralement des examinateurs ayant très peu travaillé. Considérant les résultats des 3 notes, A, B et L, les examinateurs 3F, 6F, 9H, 10F et 16H ont des valeurs d'ajustement problématiques. De ces 5 examinateurs, tous sauf 6F ont évalué moins de 100 candidats et aucun résultat concernant ces 5 examinateurs ne sera présenté dans les sections subséquentes. De plus, le total des évaluations faites par ces 5 examinateurs représente un très faible pourcentage de toutes les évaluations, ce qui amoindrit grandement le risque de biaiser les résultats des analyses subséquentes. L'examinatrice 6F, elle, est retenue parce que ses valeurs d'ajustement ne sont que légèrement trop élevées, et ce pour les

seules notes B. Les examinatrices 17F et 20F sont également retenues, car leurs valeurs inadéquates d'ajustement sont artificiellement induites par la présence des indices trop élevés des 5 examinateurs susmentionnés. Les résultats des autres analyses sont très satisfaisants et ne suggèrent aucune violation importante de l'unidimensionnalité ou de l'indépendance locale. Les valeurs estimées des niveaux de sévérité des examinateurs peuvent ainsi être utilisées pour la 2^{e} étape analytique, qui elle permettra d'atteindre les objectifs spécifiques de recherche. Cette 2^{e} étape modélisera les niveaux de sévérité des examinateurs de deux manières. La première, pour laquelle le temps représente l'expérience professionnelle, soit le nombre de candidats évalués et où chaque examinateur est étudié individuellement. C'est la section 4.2 qui suit. La seconde manière, où le temps représente le temps chronologique, permettra d'étudier l'évolution des niveaux de sévérité de plusieurs examinateurs simultanément et d'ainsi les comparer directement. Cette manière sera opérationnalisée par 5 ensembles de données distincts, représentant diverses combinaisons d'examinateurs et de périodes chronologiques et ces 5 ensembles de données seront étudiés dans les sections 4.3 à 4.7.

4.2 Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du nombre de candidats évalués

4.2.1 Représentation graphique et description

Les figures 4.1 à 4.12 montrent les graphiques chronologiques des séries chronologiques des 12 examinateurs ayant évalué au moins 100 candidats. Chaque temps de mesure représente l'évaluation de 10 candidats, et l'ordonnée de tous les graphiques est en logit. Afin de faciliter la modélisation et de permettre la comparaison visuelle des mesures de dispersion, les séries sont centrées ($m = 0$). Le 0 est arbitraire et les niveaux de sévérité des scores A, B et L ne peuvent être directement comparés, même pour un seul examinateur, puisque les valeurs obtenues des niveaux de sévérité proviennent toutes d'analyses de Rasch à multifacettes

distinctes. En revanche, les mesures de dispersion et de forme peuvent, elles, être comparées, pour un même examinateur ou d'un examinateur à l'autre. Sur chaque graphique, le trait continu représente les valeurs brutes de la série chronologique et la droite en trait pointillé représente sa tendance linéaire globale. L'ordonnée des 3 graphiques de chaque examinateur a la même étendue, ce qui permet une comparaison directe des fluctuations.

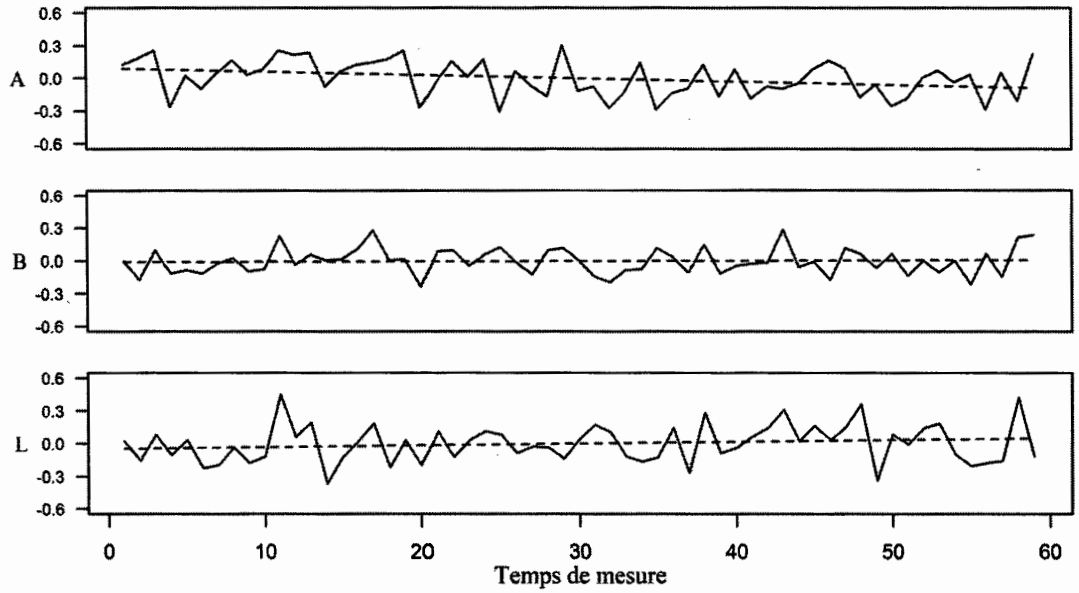


Figure 4.1 : Niveaux de sévérité de l'examinateur 1H. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

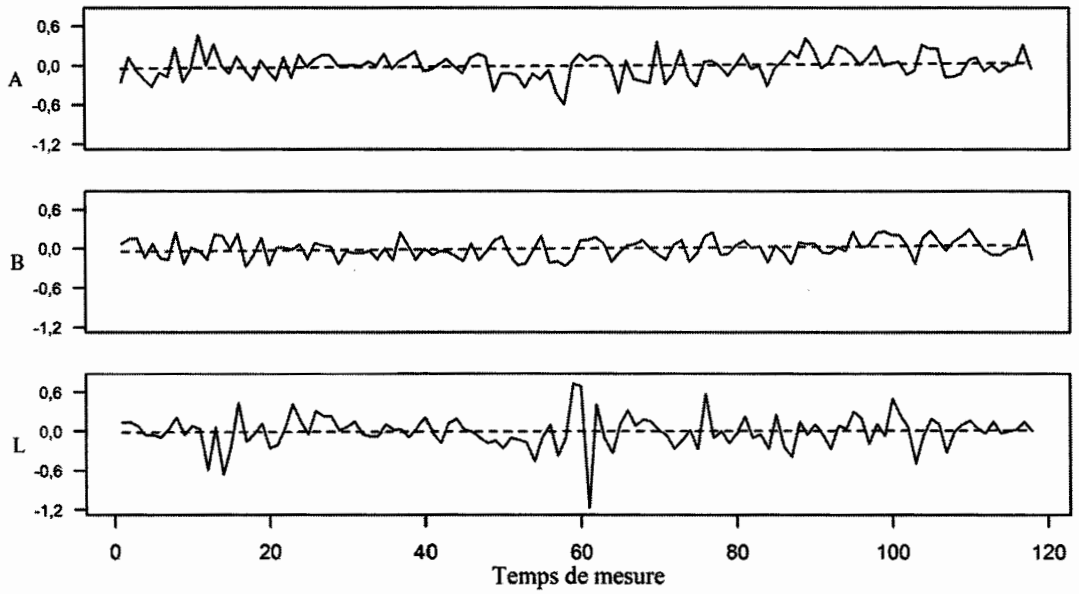


Figure 4.2 : Niveaux de sévérité de l'examinatrice 4F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

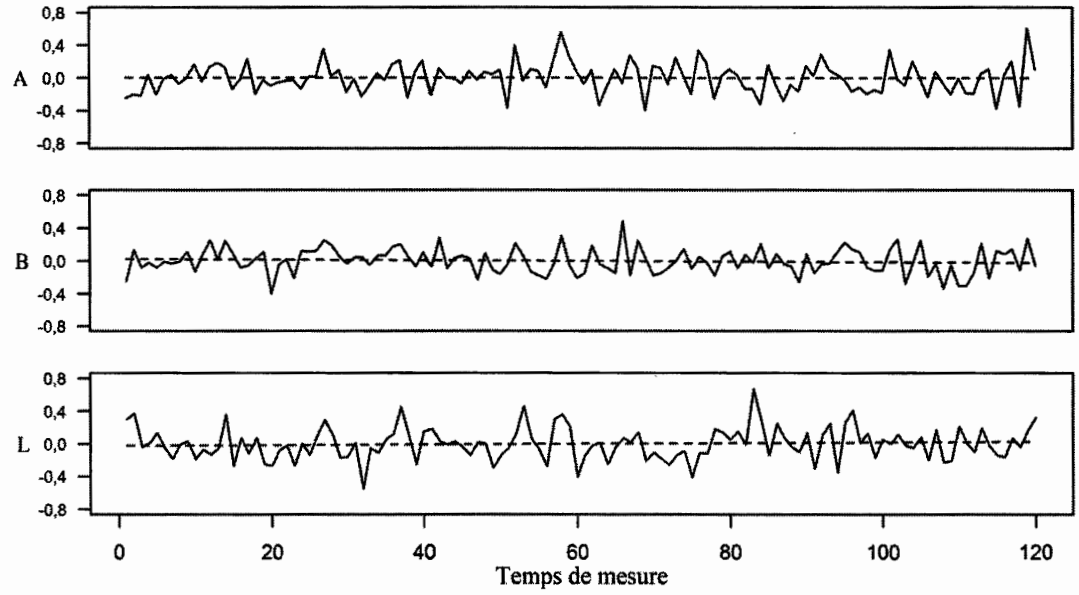


Figure 4.3 : Niveaux de sévérité de l'examinatrice 5F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

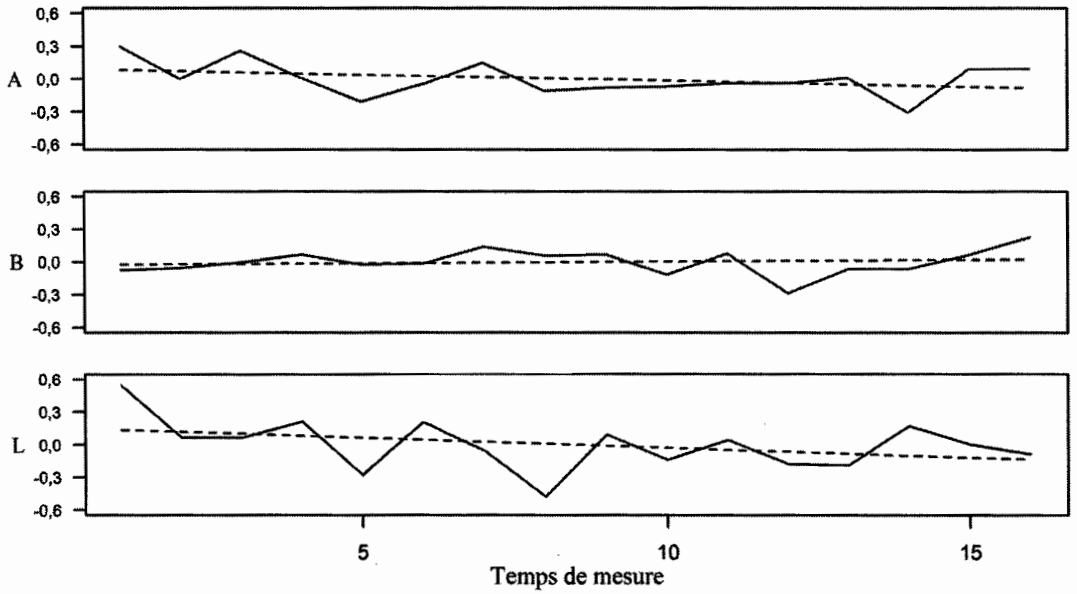


Figure 4.4 : Niveaux de sévérité de l'examinatrice 6F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

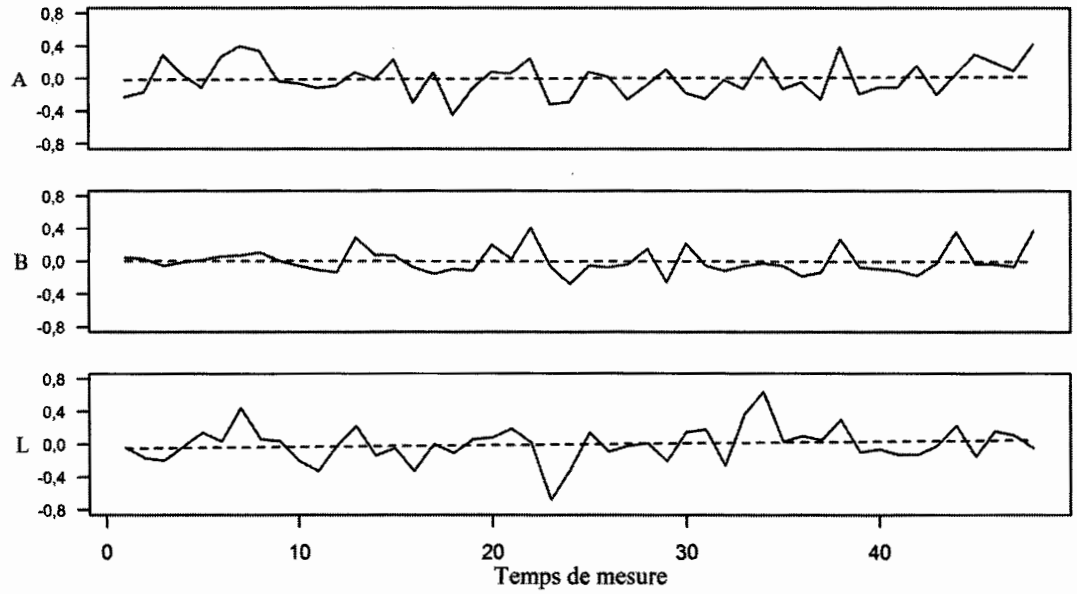


Figure 4.5 : Niveaux de sévérité de l'examinatrice 7F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

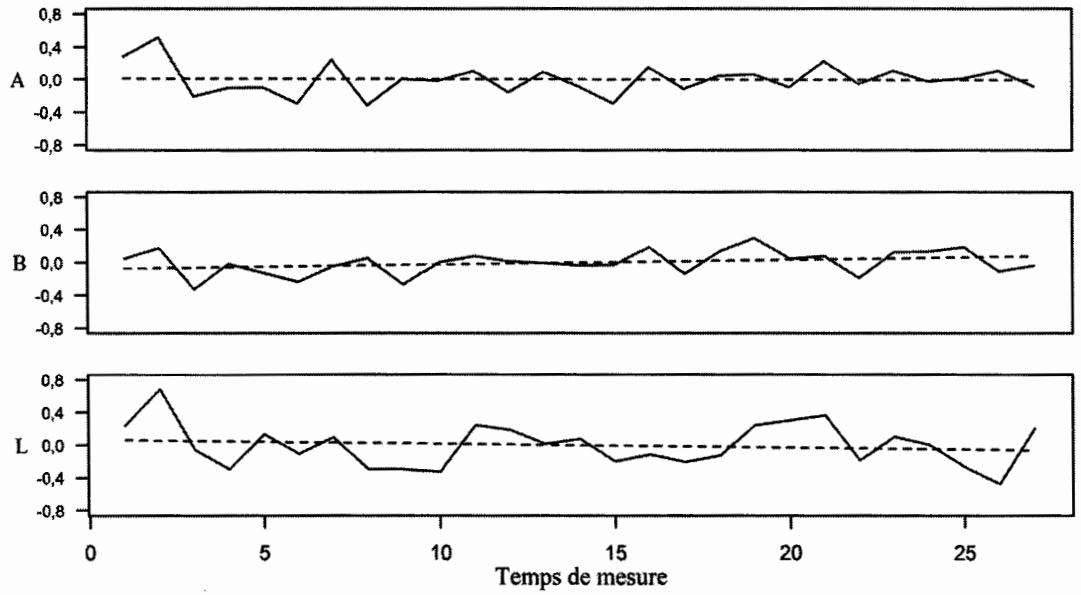


Figure 4.6 : Niveaux de sévérité de l'examinatrice 8F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

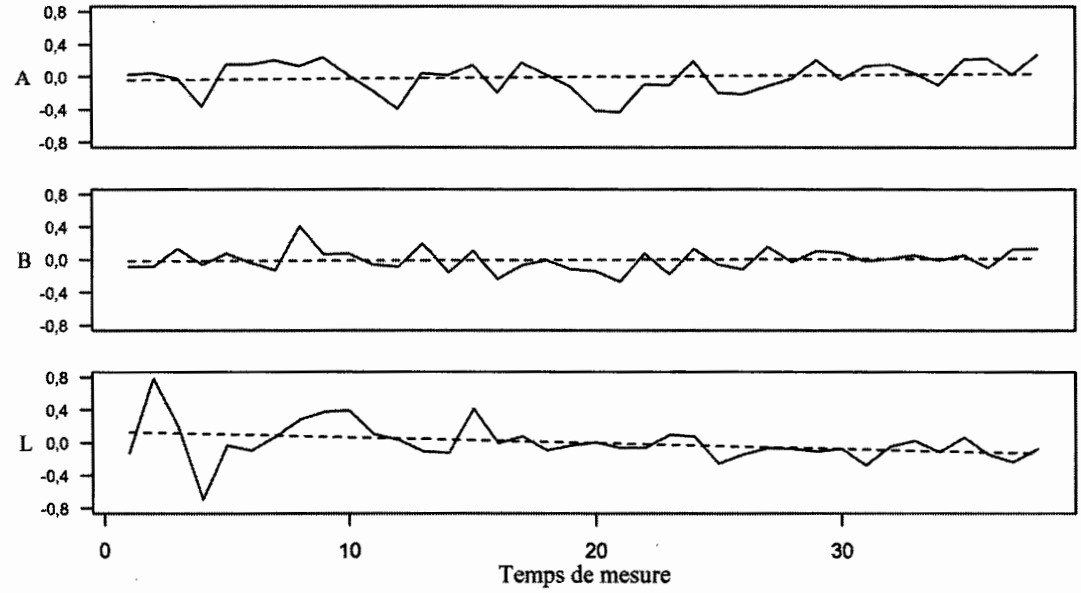


Figure 4.7 : Niveaux de sévérité de l'examinatrice 12F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

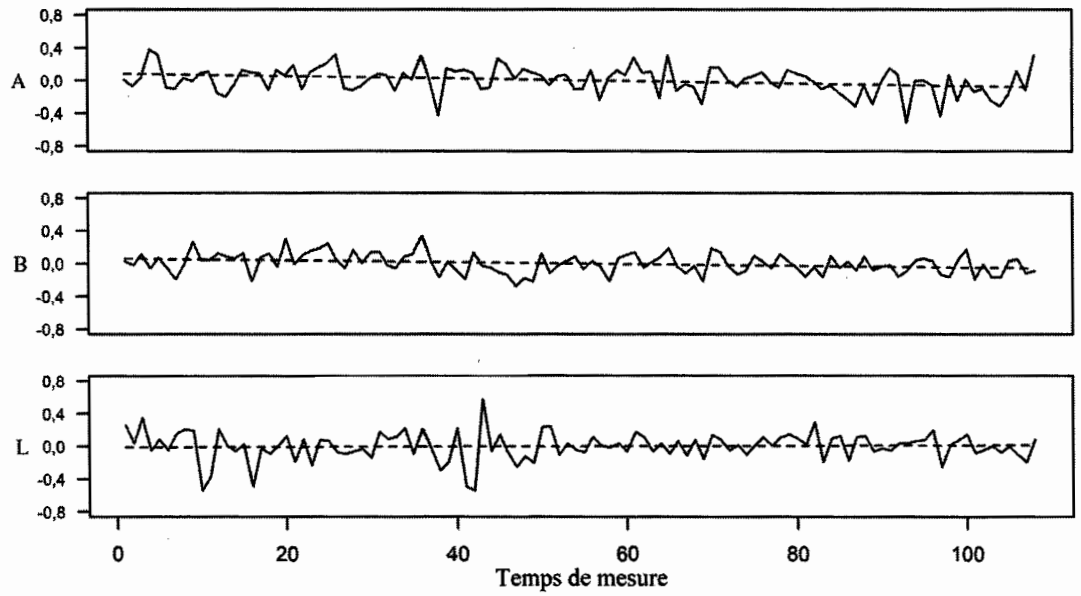


Figure 4.8 : Niveaux de sévérité de l'examinateur 13H. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

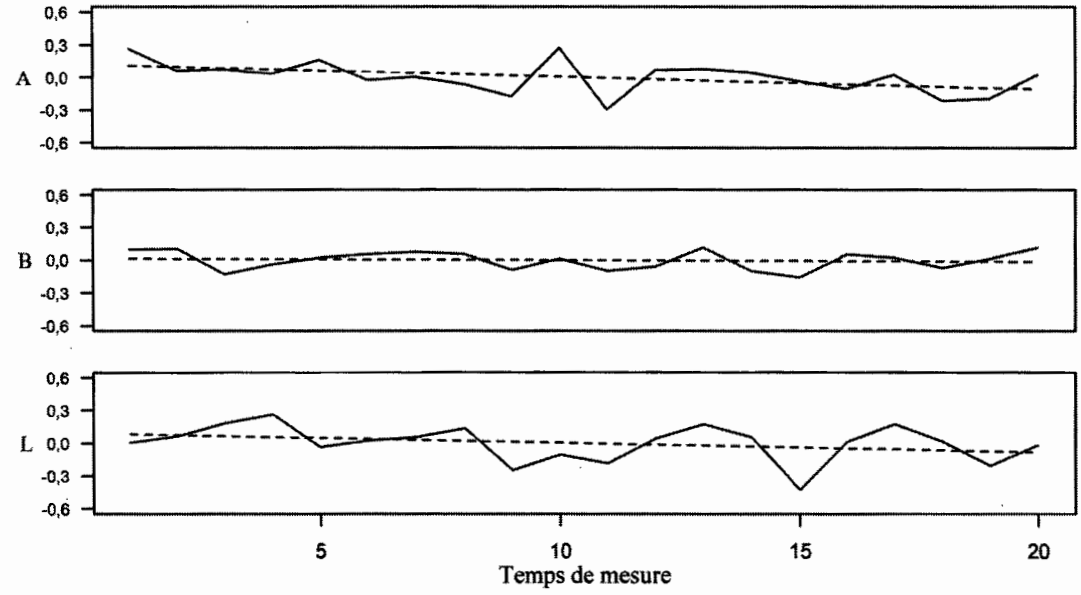


Figure 4.9 : Niveaux de sévérité de l'examinatrice 15F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

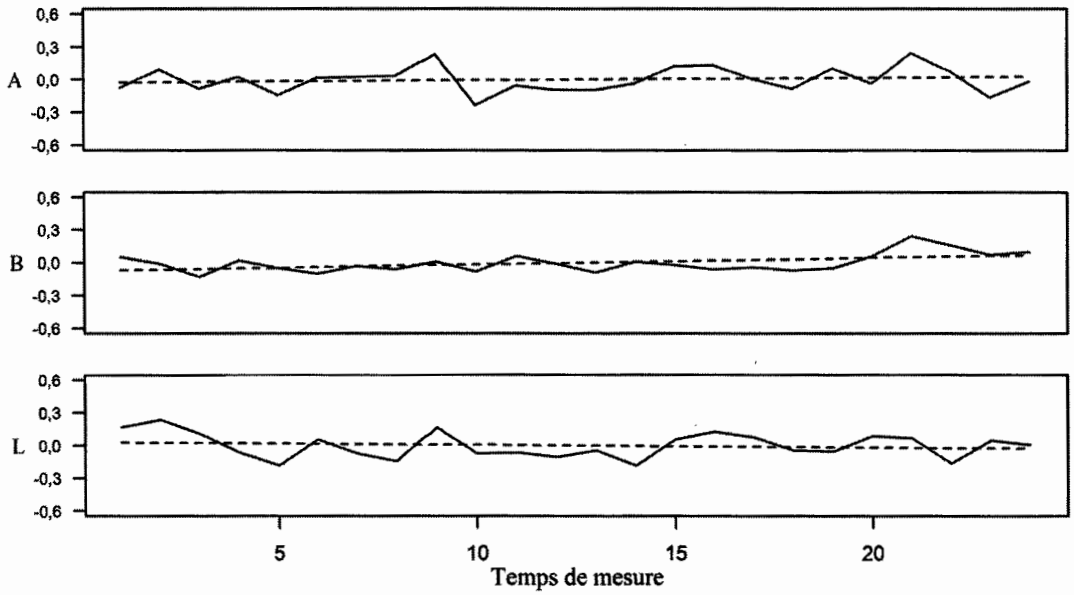


Figure 4.10 : Niveaux de sévérité de l'examinatrice 17F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

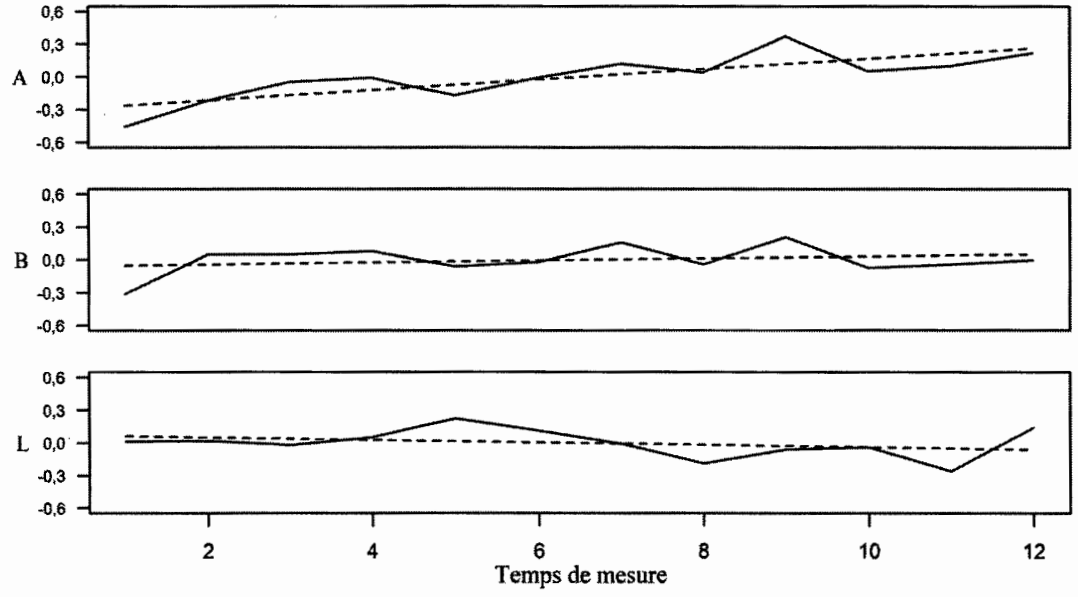


Figure 4.11 : Niveaux de sévérité de l'examinatrice 18F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

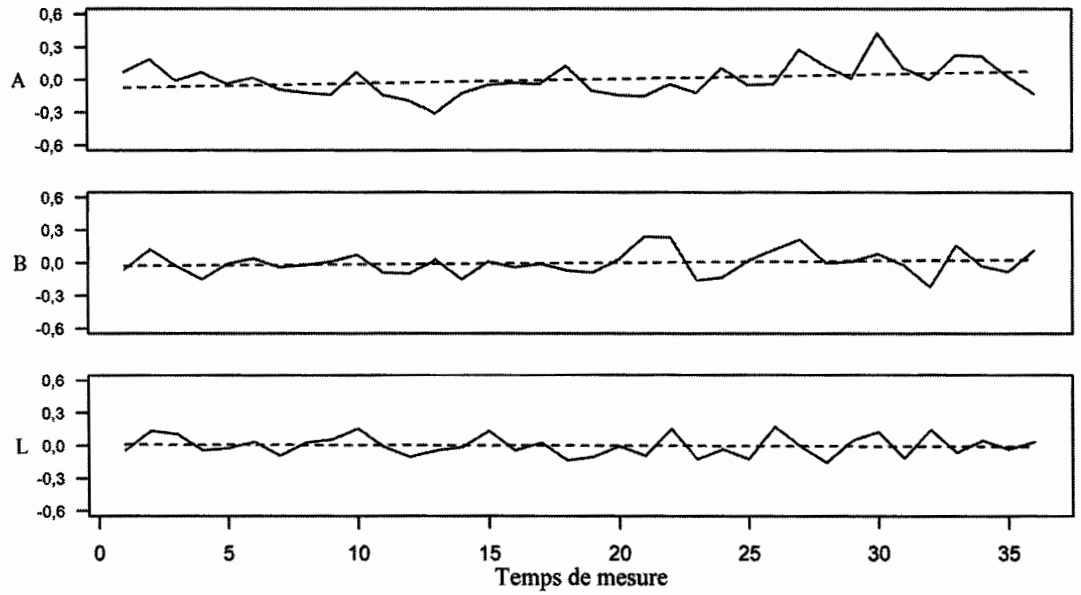


Figure 4.12 : Niveaux de sévérité de l'examinatrice 20F. Chaque temps de mesure équivaut à l'évaluation de 10 candidats

L'examen des graphiques chronologiques révèle quelques caractéristiques générales[17]. Les niveaux de sévérité sont, pour un examinateur donné, homogènes et il semble que, à en juger par la faiblesse des tendances linéaires globales, les examinateurs aient un niveau de sévérité relativement constant tout au long de leur période d'exercice, et ce bien que cette période s'étale sur plus de 2 ans pour certains d'entre eux (1H, 4F, 5F et 6F). Seules 4 séries chronologiques ont une tendance linéaire globale dont la pente est suffisamment prononcée pour être facilement perceptible à l'œil nu : 6F L, 12F L, 15F A et 18F A. Nous pouvons donc affirmer que les niveaux de sévérité sont stables et les écarts à la moyenne sont assez faibles, les valeurs extrêmes assez rares. Pour les trois ensembles de données des 12 examinateurs, lorsque nous considérons l'ensemble des valeurs de niveau de sévérité pour chaque série chronologique, une moyenne de 96 % des valeurs se trouve à ±2

[17] L'étendue en logit de chaque série semble faible, mais cela est dû à l'échelle utilisée par les données brutes, qui va de 0 à 20. Si, par exemple, les données brutes avaient été réduites de 0 à 20 au nombre de niveaux de performance, soit de 0 à 6, l'étendue en logit aurait été environ 3 fois plus grande.

écarts types de la moyenne, ce qui est très près du pourcentage de valeurs à ±2 écarts types d'une distribution normale (95,45 %). Les différences importantes entre 2 valeurs consécutives sont aussi assez rares. Il y a quelques exemples de différences très élevées (séries 4F L, temps 60 à 61 et 61 à 62 ; 12F L, temps 1 à 2 et 3 à 4 ; 13H L, temps 42 à 43), mais, en général, les différences entre valeurs successives sont modestes.

Les valeurs oscillent autour de la tendance linéaire globale et les plateaux, soit des périodes où toutes les valeurs consécutives sont au-dessus ou en dessous de cette tendance sont rares, bien qu'il y en ait trois exemples. Le premier cas est la série A de l'examinatrice 4F, où le niveau de sévérité des temps 49 à 58 est sous la tendance, ce qui représente tout de même l'évaluation de 100 candidats avec un niveau de sévérité plus clément. Les deux autres cas sont les temps 46 à 55 de la série L de la même examinatrice et les temps 69 à 77 de la série L de l'examinatrice 5F. De même, les tendances linéaires locales prononcées sont rares ; si l'on subdivise les séries chronologiques ayant au moins 40 temps de mesure, par exemple en les séparant en *n* subdivision de 20 temps, les tendances linéaires de chaque *n* subdivision sont semblables. Plusieurs observations intéressantes peuvent être faites concernant la dispersion, la variance des 36 séries. La série L de chaque examinateur a très souvent la plus grande variance et la série B a, elle, toujours la plus faible variance. En accord avec cela, les minima et maxima absolus pour chaque examinateur se trouvent dans la série L, sauf pour 18F et 20F, où la série A comporte le minimum et le maximum absolus. En revanche, la répartition dans le temps des minima et maxima de chaque série semble aléatoire, certaines séries ayant un minimum et un maximum très proches (1H A, 1H L, 4F L...), d'autres très éloignés (4F B, 13H A...), sans qu'il n'y ait de structure apparente. Des 36 séries, 34 semblent homoscédastiques, c'est-à-dire que la variance semble à peu près constante pour l'ensemble de la série, sans que des sections étendues aient une variance visiblement différente du reste de la série. Les

séries 12F L et 13F L semblent plutôt hétéroscédastiques, en ce que leur variance diminue au fil du temps, la première moitié de chaque série ayant une variance clairement supérieure à la seconde moitié. Certaines séries témoignent d'un ajustement initial plus ou moins important du niveau de sévérité, c'est-à-dire des séries où la ou les quelques premières valeurs sont éloignées de la tendance linéaire globale. Les séries 6F L, 8F A et L ainsi que 12F L ont toutes leur maximum dans les deux premiers temps de mesure, suivi de valeurs proches de la tendance linéaire globale, sauf dans le cas de la série 12F L, où le maximum est suivi, deux temps plus loin, du minimum de la série, pour ensuite rejoindre la tendance linéaire globale.

Finalement, les figures 4.13 à 4.15 montrent les diagrammes quantile-quantile, afin d'évaluer la normalité de la distribution des valeurs de niveau de sévérité lorsque ces derniers sont considérés sans égard à l'ordre chronologique. Le tableau 4.13, lui, montre les statistiques descriptives des données des 36 séries.

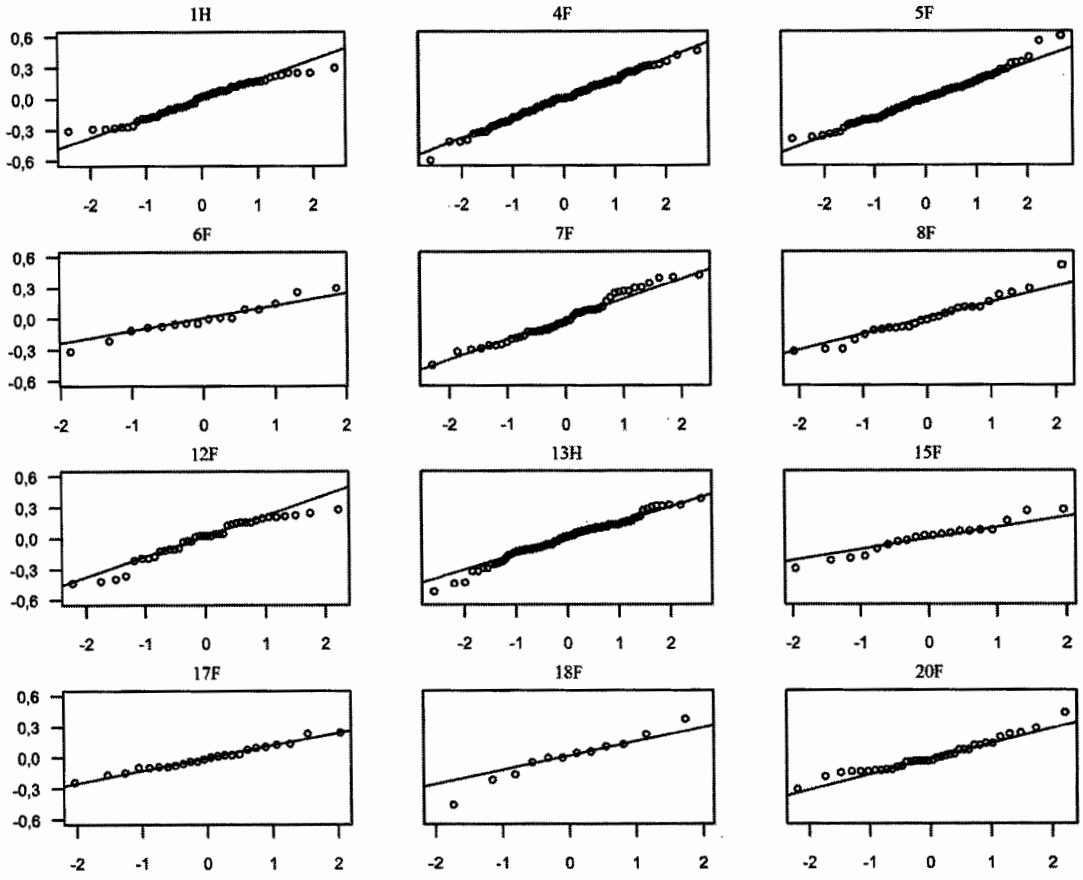


Figure 4.13 : Diagrammes quantile-quantile des séries chronologiques A de 12 examinateurs en fonction du nombre de candidats évalués

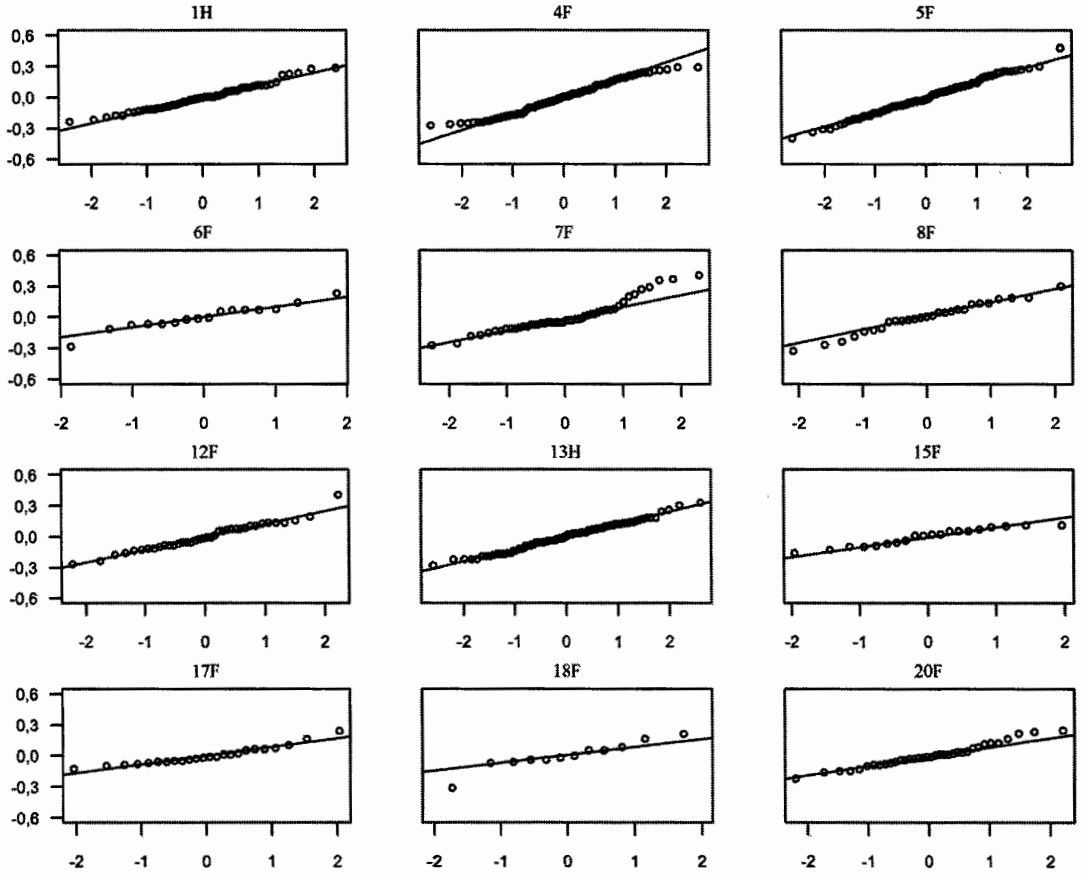


Figure 4.14 : Diagrammes quantile-quantile des séries chronologiques B de 12 examinateurs en fonction du nombre de candidats évalués

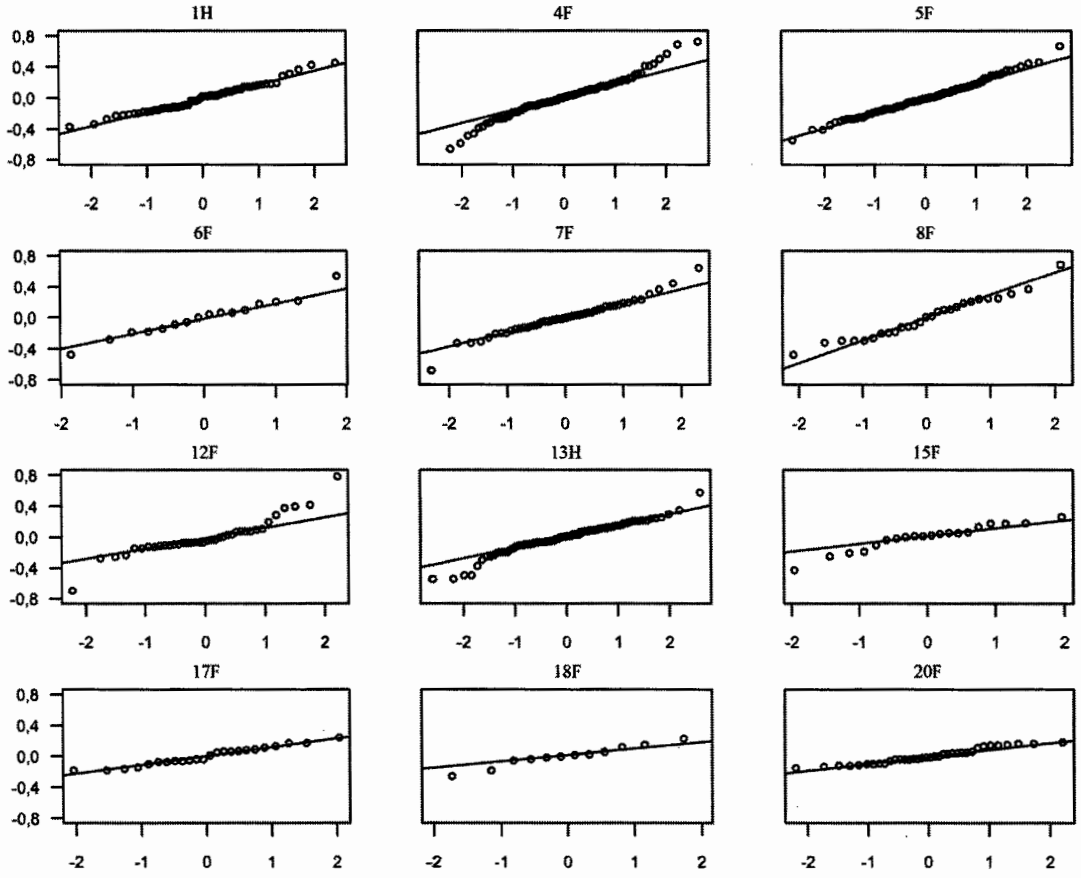


Figure 4.15 : Diagrammes quantile-quantile des séries chronologiques L de 12 examinateurs en fonction du nombre de candidats évalués

Tableau 4.13
Statistiques descriptives des 3 séries chronologiques de chacun des 12 examinateurs en fonction du nombre de candidats évalués

		t	*s*	min	max	étendue	asymétrie	aplatissement	% ±2 *s*
	A		0,17	-0,31	0,30	0,61	-0,16	-1,07	100
1H	B	59	0,10	-0,23	0,29	0,52	0,35	-0,32	97
	L		0,17	-0,37	0,45	0,82	0,37	-0,19	93
	A		0,20	-0,60	0,46	1,06	-0,22	0,05	95
4F	B	118	0,14	-0,27	0,29	0,56	0,05	-0,94	100
	L		0,24	-1,17	0,73	1,90	-0,63	3,90	95
	A		0,17	-0,40	0,61	1,01	0,37	0,48	96
5F	B	120	0,14	-0,40	0,48	0,88	0,08	-0,04	98
	L		0,20	-0,55	0,67	1,22	0,32	0,50	94
	A		0,14	-0,31	0,30	0,61	0,13	-0,40	100
6F	B	16	0,10	-0,28	0,23	0,51	-0,39	0,35	94
	L		0,24	-0,48	0,54	1,02	0,19	0,10	88
	A		0,22	-0,45	0,42	0,87	0,23	-0,83	98
7F	B	48	0,14	-0,27	0,41	0,68	0,96	5,30	94
	L		0,22	-0,68	0,64	1,32	0,01	1,59	94
	A		0,20	-0,32	0,51	0,83	0,50	0,24	96
8F	B	27	0,14	-0,33	0,29	0,62	-0,31	-0,48	96
	L		0,26	-0,47	0,69	1,16	0,41	-0,28	96
	A		0,20	-0,43	0,28	0,71	-0,64	-0,49	92
12 F	B	38	0,10	-0,27	0,40	0,67	0,46	0,63	95
	L		0,22	-0,69	0,79	1,48	0,61	2,89	95
	A		0,17	-0,52	0,38	0,90	-0,40	0,43	96
13 H	B	108	0,10	-0,28	0,33	0,61	0,10	-0,31	96
	L		0,17	-0,54	0,58	1,12	-0,55	1,80	94
	A		0,14	-0,30	0,27	0,57	-0,08	-0,53	100
15 F	B	20	0,10	-0,16	0,11	0,27	-0,27	-1,35	100
	L		0,17	-0,43	0,27	0,70	-0,75	0,10	95
	A		0,10	-0,24	0,25	0,49	0,26	-0,49	96
17 F	B	24	0,10	-0,13	0,24	0,37	0,96	0,61	96
	L		0,10	-0,18	0,24	0,42	0,13	-1,06	96
	A		0,20	-0,46	0,37	0,83	-0,39	-0,27	92
18 F	B	12	0,14	-0,31	0,21	0,52	-0,62	0,37	92
	L		0,14	-0,26	0,22	0,48	-0,30	-0,64	100
	A		0,14	-0,31	0,43	0,74	0,65	0,39	94
20 F	B	36	0,10	-0,22	0,24	0,46	0,38	-0,30	92
	L		0,10	-0,15	0,18	0,33	0,29	-1,08	100

Sur les 36 séries chronologiques, 29 sont à peu près normalement distribuées et 7 sont plus ou moins anormalement distribuées. Il s'agit des séries 1H A, 4F L, 7F B et L, 12F A et L ainsi que 13H L. Pour toutes ces séries sauf une (1H A), le problème est que la série est leptokurtique, les queues sont donc plus épaisses et les valeurs les plus élevées ou plus faibles ne suivent pas une distribution conforme à la loin normale. La série 1H A, elle, est platykurtique, et toutes ses valeurs sont situées à ±2 écarts types de la moyenne, tandis qu'une distribution normale ayant le même nombre de données (59) devrait avoir 2 ou 3 valeurs situées à plus de 2 écarts types de la moyenne. Il est toutefois intéressant de constater que les 36 séries chronologiques sont assez symétriques : 28 des 36 séries ont un coefficient d'asymétrie compris entre -0,5 et +0,5 et seules 3 séries ont un coefficient égal ou supérieur à 0,75 en valeur absolue. Cela montre que chaque examinateur a une tendance à peu près égale, par rapport à son niveau de sévérité moyen, à être un peu plus sévère ou un peu plus clément, sans préférence nette pour l'un ou l'autre.

4.2.2 Modélisation AMMI

Des 36 séries chronologiques modélisées, 24 ne sont statistiquement que du bruit blanc, c'est-à-dire que le modèle AMMI décrivant le mieux la série est le modèle 0,0,0. Rappelons que la modélisation AMMI a 3 paramètres. Le premier est le paramètre autorégressif (« A ») qui décrit le lien entre les valeurs d'une série chronologique au temps t et $t - p$, où p égale l'ordre du paramètre. Par exemple, un paramètre A d'ordre 2 indique une covariance entre les valeurs d'une série au temps t et $t - 2$. Le deuxième paramètre indique l'ordre d'intégration (« I »), c'est-à-dire le nombre de fois qu'une série chronologique doit être différenciée afin de respecter les conditions d'utilisation de la modélisation AMMI. Le troisième paramètre représente le paramètre de moyenne mobile (« MM »), soit la covariance entre la valeur de la série au temps t et l'erreur au temps $t - q$, où q représente l'ordre de ce paramètre. Ainsi, un modèle 0,0,0 représente le modèle nul, où la valeur d'une série

chronologique au temps t égale simplement la moyenne générale de la série à laquelle s'ajoute une erreur locale. Les modèles le mieux ajustés aux données des 12 autres séries sont présentés dans le tableau 4.14.

Tableau 4.14
Modèles AMMI des séries chronologiques en fonction du nombre de candidats évalués

Série	AMMI	Coefficients*	Erreurs types
1H A	0,1,1	-0,90	0,06
4F A	1,0,0	0,18	0,09
4F B	1,1,2	-0,74 ; -0,03 ; -0,87	0,11 ; 0,07 ; 0,06
5F A	3,0,2	-0,65 ; -0,74 ; 0,21 ; 0,66 ; 0,81	0,17 ; 0,14 ; 0,11 ; 0,17 ; 0,08
12F A	1,0,0	0,30	0,16
12F B	5,0,0	-0,12 ; 0,17 ; 0,01 ; -0,04 ; 0,46	0,14 ; 0,15 ; 0,15 ; 0,15 ; 0,14
12F L	0,0,2	0,29 ; -0,29	0,15 ; 0,15
13H A	0,1,1	-0,93	0,03
13H B	0,1,1	-0,90	0,06
15F A	0,1,2	-1,05 ; 0,62	0,17 ; 0,46
17F B	0,1,1	-0,53	0,23
20F A	0,1,1	-0,66	0,15

*Les coefficients suivent l'ordre $A_1...A_p$, $MM_1...MM_q$ et sont séparés par un « ; »

Tous les modèles respectent les conditions d'utilisation de la modélisation AMMI, à l'exception des séries 12F A et 20F A, qui semblent hétéroscédastiques (voir les résultats des tests diagnostiques en annexe A). Pour ces 2 séries, l'autocorrélation des résidus au carré est élevée au délai 2 (12F A) ou 3 (20F A), ce qui est l'un des signes d'une possible hétéroscédasticité, et les résultats de certains tests diagnostiques révèlent de potentiels problèmes d'hétéroscédasticité. La modélisation AMMI a néanmoins été utilisée pour ces 2 séries chronologiques, et ce pour trois raisons. Premièrement, l'inspection visuelle des graphiques chronologiques montre que l'hétéroscédasticité n'est pas trop importante. Deuxièmement, les autres conditions

d'utilisation semblent respectées et, troisièmement, les modèles pour séries hétéroscédastiques (GARCH) demandent des tailles d'échantillon de plusieurs centaines de temps de mesure (Hwang et Valls Pereira, 2006 ; Zumbach, 2000).

Les coefficients du tableau 4.14 s'interprètent comme des coefficients de régression (ce qu'ils sont, en fait). Plus un coefficient a une valeur absolue élevée, plus forte est la dépendance entre la valeur du niveau de sévérité au temps $t - k$ et la valeur du niveau de sévérité au temps t. À noter que, pour les modèles ayant un ordre supérieur à 1, seul le coefficient de l'ordre le plus élevé est d'intérêt. Par exemple, pour la série 12F B, seul le coefficient d'ordre 5 importe. Des 12 modèles du tableau 4.14, un seul concerne une série chronologique L (celle de 12F) et son coefficient d'ordre 2 est de petite taille : c'est le 3^{e} plus faible coefficient du tableau pour les coefficients d'ordre le plus élevé. Il est néanmoins difficile de voir si la rareté des séries L ayant un modèle non nul est le fruit du hasard ou un résultat substantiellement intéressant. L'autre série particulière est la série 12F B, dont le modèle est d'ordre 5, ce qui est un ordre très élevé que l'on retrouve rarement en sciences sociales ou en psychologie (Shin, 2017 ; Snippe, Bos, van der Ploeg, Sanderman, Fleer et Schroevers, 2015). Dans ce cas, cela signifierait que le niveau de sévérité de 12F, lors de l'évaluation des candidats 51 à 60 pour les critères d'évaluations B, serait lié au niveau de sévérité qu'elle avait lors de l'évaluation des candidats 1 à 10, et ainsi de suite. Il est très difficile de supposer un mécanisme psychologique ou environnemental plausible pour une telle relation ; il s'agit probablement d'un « faux positif », d'autant plus que si le coefficient d'ordre 5 est d'une taille modérée (0,46), les 4 coefficients d'ordre inférieur sont tous très faibles (de 0,01 à 0,17) et laissent croire que le coefficient significatif d'ordre 5 est le fruit du hasard. Par ailleurs, 3 examinateurs ont un modèle non nul pour plus d'une série chronologique : 4F, 12F et 13H. Les modèles de ces deux derniers examinateurs ont la particularité d'avoir des coefficients de taille similaire, en valeur absolue, pour leurs multiples séries, soit de tailles moyennes pour

12F (de 0,29 à 0,46) et de grandes tailles pour 13H (0,90 et 0,93). Ce résultat est intéressant, car il est compatible avec l'idée de mécanismes psychologiques régissant le niveau de sévérité d'un examinateur donné, mais il ne faut pas surinterpréter ce résultat. Quoi qu'il en soit, il y a plusieurs modèles ayant des coefficients très élevés, ce qui montre qu'une forte dépendance entre les valeurs du niveau de sévérité existe pour certains examinateurs (1H A, 4FB, 13HA et B). Pour terminer, relevons que, bien que les modèles les mieux ajustés de 7 séries sur 12 soient différenciés, ces séries ont une tendance linéaire globale plutôt faible. Le tableau 4.15 présente les tendances linéaires globales des 7 séries différenciées et la différence en logit correspondante entre le 1er et le dernier temps de mesure. Étant donné les très faibles valeurs, les résultats sont présentés à la 4^{e} décimale.

Tableau 4.15
Pentes de la tendance linéaire des séries chronologiques différenciées

	AMMI	Pente de la tendance	Erreurs types	Différence en logit
1H A	0,1,1	-0,0030	0,0012	-0,18
4F B	1,1,2	0,0009	0,0004	0,10
13H A	0,1,1	-0,0016	0,0050	-0,17
13H B	0,1,1	-0,0011	0,0004	-0,12
15F A	0,1,2	-0,0116	0,0052	-0,22
17F B	0,1,1	0,0060	0,0023	0,14
20F A	0,1,1	0,0043	0,0023	0,15

Les écarts entre les niveaux moyens au premier et au dernier temps sont minimes, puisque 0,15 logit, soit l'écart moyen pour ces sept séries, n'a qu'un faible impact sur les notes accordées par un examinateur. En fait, l'écart le plus important observé chez ces 12 examinateurs ne se trouve pas dans le tableau 4.15, puisqu'il appartient à la série A de l'examinatrice 18F. Comme il n'y a que 12 temps de mesure, le modèle de cette série n'est pas différencié en dépit de l'écart de 0,53 logit entre le premier et le

dernier temps de mesure – par manque de puissance statistique. À l'exception de cette série, très courte, le niveau de sévérité de ces 12 examinateurs est temporellement stable, du moins d'un point de vue pragmatique.

4.2.3 Corrélations croisées intraindividuelles

L'un des aspects de la modélisation de l'évolution du niveau de sévérité des examinateurs concerne les liens entre les niveaux de sévérité A, B et L pour chacun des examinateurs. C'est un aspect important, puisque d'éventuels liens entre les 3 niveaux de sévérité seraient un élément de preuve quant à l'existence d'un niveau de sévérité global qui serait un trait psychologique, une propriété de l'individu. Les corrélations croisées intraindividuelles ont donc été calculées pour chacun des 12 examinateurs étudiés dans cette section, aux délais -1, 0 et +1[18]. La figure 4.16 montre la distribution de 12 corrélations croisées intraindividuelles entre les séries A, B et L, aux délais -1, 0 et +1.

[18] Dans ce texte, la convention utilisée est la suivante : une corrélation croisée entre x et y au délai -1 renvoie à la corrélation entre x au temps $t-1$ et y au temps t. Inversement, le délai +1 réfère à la corrélation entre y au temps $t-1$ et x au temps t.

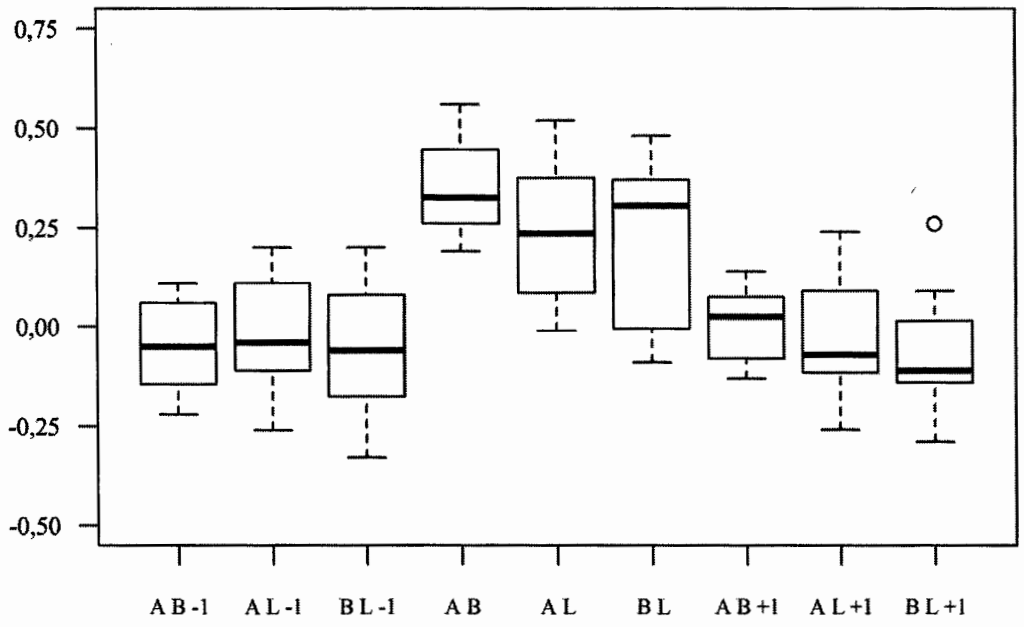


Figure 4.16 : Corrélations croisées aux délais -1, 0 et +1 entre les séries A, B et L. Chaque graphique en boîte et moustaches montre la distribution de 12 coefficients de corrélations croisées

Au délai 0 (les 3 graphiques centraux), les corrélations entre A et B sont en moyenne plus élevées que les corrélations entre A et L ou B et L et cela se comprend aisément, puisque les scores A et B sont attribués à la communication, tandis que le score L ne concerne que la langue. Les corrélations entre A et B sont assez homogènes, allant de 0,19 à 0,56, plus que les corrélations entre A et L ou B et L, dont l'étendue est environ deux fois plus grande. Les corrélations aux délais -1 et +1 sont plus ou moins symétriquement distribuées autour de 0, et elles sont clairement différentes des corrélations au délai 0, ce qui montre que, pour ces 12 examinateurs, le niveau de sévérité pour un score au temps t ne mène pas, n'a pas de lien avec le niveau de sévérité des autres scores aux temps $t - 1$ ou $t + 1$. Les niveaux de sévérité sont toutefois positivement corrélés au temps t, mais assez faiblement. Pour conclure, ces résultats sont compatibles avec l'idée selon laquelle il n'y a pas un seul « niveau de sévérité » assimilable à un trait de personnalité pour chaque examinateur. Si c'était le cas et que chaque examinateur avait effectivement, comme trait psychologique, un

niveau de sévérité « unique » comme le pensait DeCoths (1977) ou Guilford (1954), alors les corrélations entre les 3 niveaux de sévérité étudiés ici (A, B et L) seraient très élevées, puisque les 3 niveaux ne seraient que l'expression légèrement distincte d'un seul niveau de sévérité « global », propre à un examinateur donné. Ce n'est manifestement pas le cas ici. Il semble que le degré de similarité entre les objets évalués soit en lien avec les niveaux de sévérité propres à l'évaluation de chacun de ces objets. Dans notre cas, les niveaux de sévérité A et B sont plus fortement corrélés que les niveaux A et L ou B et L parce que A et B sont similaires et diffèrent sensiblement de L, les deux premiers concernant l'évaluation des habiletés communicatives et le dernier les habiletés linguistiques.

4.2.4 Comparaisons débutants et expérimentés

Cette section présente les résultats en lien avec le deuxième objectif spécifique de recherche : « comparer l'évolution du niveau de sévérité des examinateurs débutants et expérimentés ». Les données de cet ensemble ne permettent pas de comparer des examinateurs entre eux, puisque chaque examinateur est étudié de manière isolée, en prenant en compte l'ensemble des candidats que chaque examinateur a évalué d'octobre 2010 à avril 2014. Ces données permettent toutefois de comparer chaque examinateur à lui-même, afin de voir si le niveau de sévérité évolue de manière particulière au début de la carrière d'un examinateur. Les 12 examinateurs étudiés dans cette section comprennent 9 examinateurs débutants ayant commencé à travailler comme examinateur en octobre 2010 ou plus tard. Les 3 autres examinateurs travaillaient déjà avant que ne commence la période de collecte des données utilisées par cette thèse (1H) ou travaillaient dans un autre centre TEF avant de se joindre au centre d'où proviennent les données de cette thèse (12F et 20F). L'étude de l'évolution temporelle du niveau de sévérité des examinateurs débutants a été faite en découpant le temps en périodes de 5 candidats évalués, afin d'obtenir une granularité plus fine permettant d'étudier d'éventuelles fluctuations du niveau de

sévérité en tout début de carrière. Plus spécifiquement, les 20 premiers temps de mesure ont été isolés et la série chronologique résultante a été comparée à sa série globale. L'étude du niveau de sévérité des examinateurs débutants tient donc compte de l'évaluation des 100 premiers candidats de la carrière d'un examinateur.

Afin de gagner de l'espace, seuls les résultats indiquant la présence d'éléments particuliers lors du début de la carrière d'un examinateur seront présentés. Sur les 9 examinateurs débutants, une seule examinatrice, 4F, ne présente aucun élément notable lors des 20 premiers temps de ses 3 séries chronologiques. Les 8 autres examinateurs ont des éléments notables de 3 types : des différences entre la tendance linéaire locale des 20 premiers temps de mesure et la tendance linéaire globale de la série chronologique complète, des différences de variance entre la série des 20 premiers temps et la série complète et la présence du maximum ou du minimum de la série globale lors des 20 premiers temps de mesure. Le tableau 4.16 présente les éléments notables recensés. La colonne « Diff. tendance » représente la différence, en logit, entre les valeurs initiales et finales de la tendance linéaire du niveau de sévérité de la série chronologique complète et celle des 20 premiers temps de mesure. Par exemple, la série 6F A complète a une différence de -0,19 logit entre le premier et le dernier temps de mesure, et la série 6F A des 20 premiers temps de mesure a une différence de -0,28 logit. La différence entre les 2 est de 0,09 logit, et comme la tendance linéaire de la série des 20 premiers temps a une pente plus négative, la valeur de la différence est négative. La colonne « % de variance » indique le pourcentage de la variance de la série globale que représente la variance de la série des 20 premiers temps. Une valeur inférieure à 100 indique que la variance de la série des 20 premiers temps est inférieure à la variance de la série globale et inversement.

Tableau 4.16
Présence d'éléments notables dans les 20 premiers temps des séries chronologiques des examinateurs 5F, 6F, 7F, 8F, 13H, 15F, 17F et 18F

Série	Diff. tendance	% de variance	Minimum	Maximum
5F A	0,38	45	Non	Non
5F L	-0,43	67	Non	Non
6F A	-0,09	111	Non	Oui
6F L	-0,26	127	Oui	Oui
7F A	0,26	106	Oui	Non
8F A	-0,35	183	Oui	Oui
8F B	0,00	122	Oui	Non
8F L	-0,53	131	Non	Oui
13H A	0,00	138	Non	Oui
13H L	-0,15	109	Non	Non
15F L	0,00	67	Non	Oui
17F A	0,00	108	Non	Oui
17F L	-0,22	156	Oui	Oui
18F A*	0,21	113	Oui	Non
18F B*	0,24	157	Oui	Non
18F L*	0,30	67	Non	Non

* Ces séries n'ont que 10 temps de mesure, étant donné le faible nombre de temps de mesure de la série globale

Certaines de ces séries ne se différencient guère de la série globale dont elles sont issues. Les séries 8F B, 15F L et 17F A n'ont de remarquable que la seule présence de la valeur minimale ou maximale, ce qui peut très bien être le fruit du hasard. Après tout, la série 8F B globale a 55 temps de mesure. Les 20 premiers temps de mesure représentent ainsi 36 % de la série totale et le fait d'y trouver le minimum n'est pas particulièrement étrange. Plus intéressantes sont les séries où se trouvent plusieurs éléments notables, soit les séries 6F L, 8F A et L, 17F L ainsi que 18F A et B. Ces séries ont à la fois une plus grande variance que leur série globale respective, une

tendance linéaire locale marquée et un minimum ou un maximum global. Elles suggèrent fortement un processus d'ajustement du niveau de sévérité de ces examinatrices, suggestion renforcée par le fait que certaines examinatrices ont des éléments notables pour 2 ou 3 de leurs séries. Les examinatrices 8F et 18F ont des éléments notables pour chacune de leurs 3 séries et tous les indices concordent dans la démonstration d'une plus grande volatilité de leur niveau de sévérité lors de l'évaluation de leurs 100 premiers candidats évalués. La variance de certaines de leur série des 20 premiers temps est particulièrement importante (183 % pour 8F A et 157 % pour 18F B). Ce sont d'ailleurs les 2 seules examinatrices ayant au moins un élément notable dans leur série B.

Au total, 6 des 9 examinateurs débutants ont des éléments notables dans au moins 2 de leurs 3 séries des 20 premiers temps de mesure, ce qui montre qu'il y a, pour certains examinateurs, un processus d'ajustement intraindividuel du niveau de sévérité lors de l'insertion professionnelle comme examinateur. Ce résultat ne permet toutefois d'atteindre que partiellement notre 2^{e} objectif spécifique de recherche, car il ne permet pas une réelle comparaison synchronique du niveau de sévérité d'examinateurs débutants et expérimentés, comparaison qui sera possible lors de l'étude des autres ensembles de données dans les sections suivantes.

4.3 Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du temps chronologique, du 2010-10 au 2013-03

4.3.1 Représentation graphique et description

Les figures 4.17 à 4.19 montrent les graphiques chronologiques des séries chronologiques des examinateurs 1H et 4F, du 2010-10 au 2013-03. Chaque temps de mesure représente une période de 3 mois et il y a 10 temps de mesure. Au cours de cette période, 1H a évalué 591 candidats et 4F en a évalué 661, pour des moyennes respectives de candidats évalués par temps de mesure de 59 et 66. Les estimations de

niveau de sévérité, pour chaque note, sont communes pour les 2 examinateurs et leurs niveaux de sévérité peuvent donc être directement comparés. Les figures montrent les niveaux de sévérité dans un même graphique. Afin de favoriser la comparaison des fluctuations entre les séries des 3 notes, l'étendue de l'ordonnée est la même pour les 3 graphiques, soit 1,20 logit. Les lignes pointillées représentent la tendance linéaire globale de chaque série.

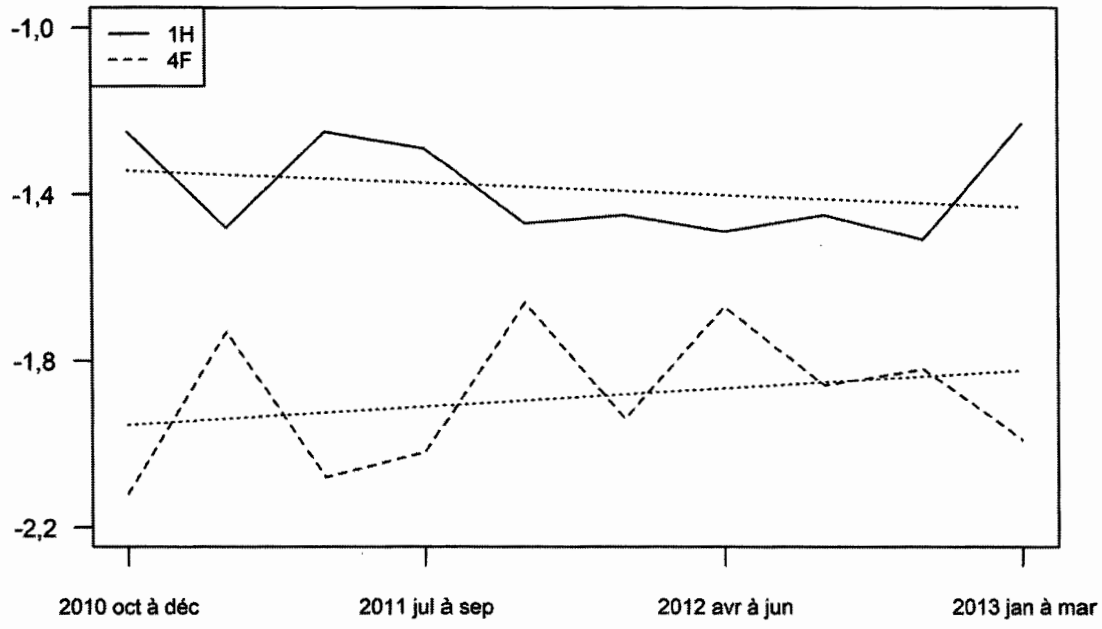


Figure 4.17 : Niveaux de sévérité A des examinateurs 1H et 4F, du 2010-10 au 2013-03. L'ordonnée est en logit.

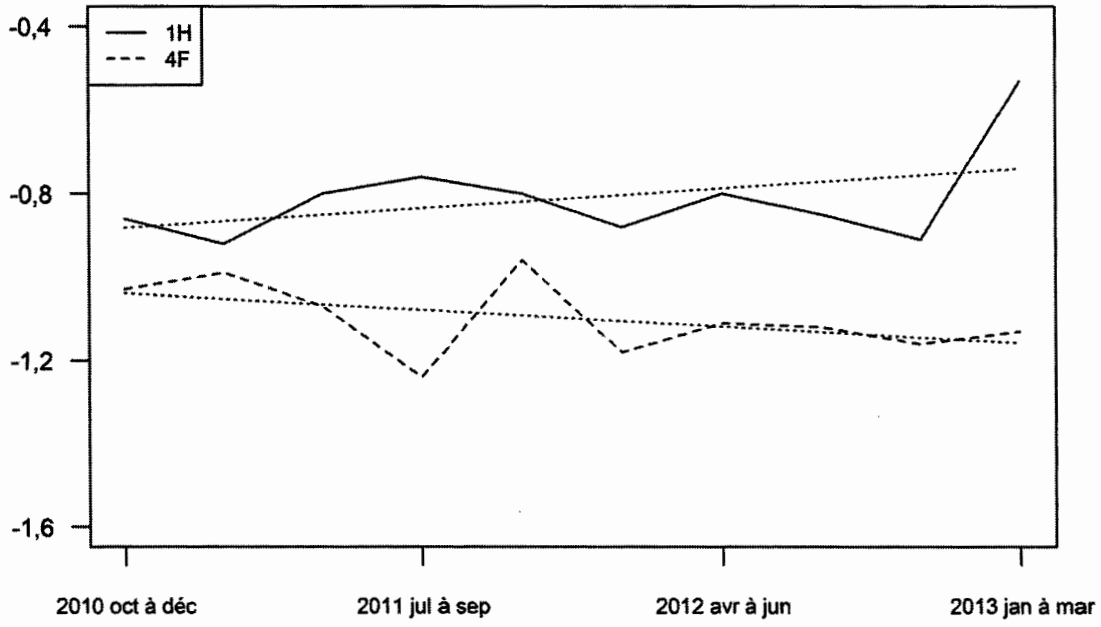


Figure 4.18 : Niveaux de sévérité B des examinateurs 1H et 4F, du 2010-10 au 2013-03. L'ordonnée est en logit.

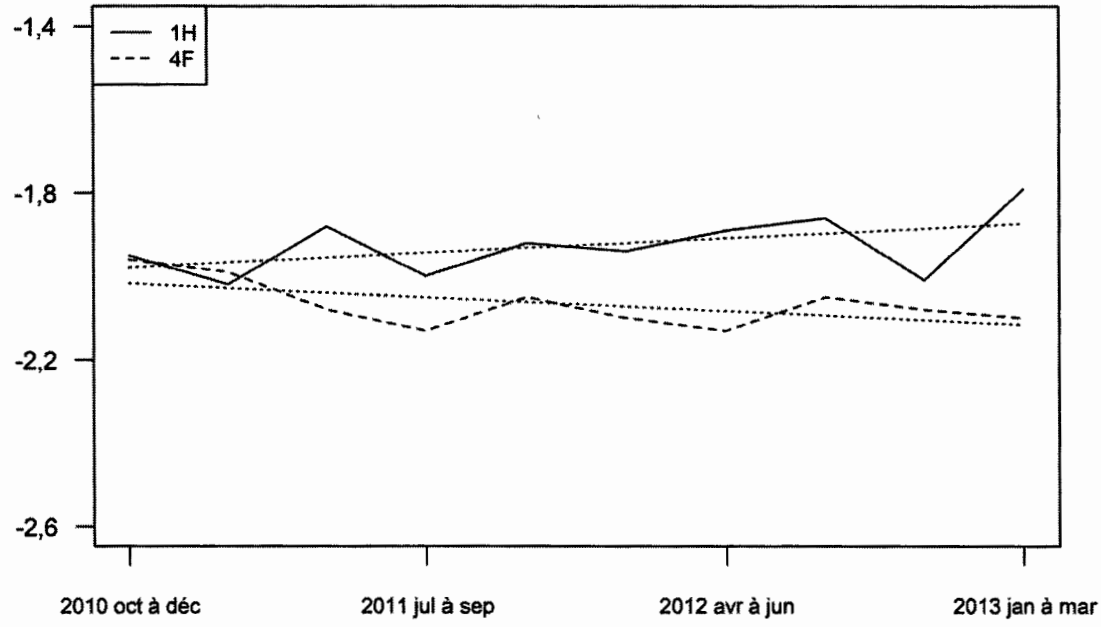


Figure 4.19 : Niveaux de sévérité L des examinateurs 1H et 4F, du 2010-10 au 2013-03. L'ordonnée est en logit.

Les niveaux de sévérité des 2 examinateurs sont stables, leur tendance linéaire globale ayant une pente très faible. Pour ces 6 séries chronologiques, l'écart maximal entre la valeur de la tendance linéaire au dernier temps et sa valeur au premier temps est de 0,14 logit, ce qui est négligeable. Il n'y a pas de valeurs extrêmes ; pour 5 des 6 séries, 100 % des valeurs sont situées à 2 écarts types de la moyenne, l'autre série (1H B) ayant une valeur à l'extérieur de ces bornes. Puisque ces séries n'ont que 10 temps de mesure, il ne saurait être question d'adéquation à une distribution statistique, normale ou autre, aussi ni les statistiques descriptives ni les diagrammes quantile-quantile ne sont présentés. Remarquons, pour terminer, que l'ordre de sévérité entre ces 2 examinateurs est presque parfaitement respecté pour les 3 séries. L'examinateur 1H est toujours plus sévère que l'examinatrice 4F, sauf au temps 2 des séries L, où le niveau de sévérité de 1H est inférieur de 0,03 au niveau de sévérité de 4F. Il s'agit du seul croisement entre les courbes des deux examinateurs.

4.3.2 Modélisation AMMI

Les 6 séries chronologiques sont du bruit blanc, le modèle 0,0,0 étant le mieux ajusté aux données. Considérant que chaque série n'a que 10 temps de mesure, cela n'est pas surprenant.

4.3.3 Corrélations croisées intraindividuelles

Le tableau 4.17 montre les corrélations croisées intraindividuelles entre les niveaux de sévérité des 3 séries chronologiques.

Tableau 4.17
Corrélations croisées intraindividuelles aux délais -1, 0 et +1 pour les examinateurs 1H et 4F, du 2010-10 au 2013-03

	Délai		
	-1	0	+1
1H : A et B	-0,23	0,61	-0,16
1H : A et L	-0,60	0,38	-0,30
1H : B et L	-0,18	0,73	-0,39
4F : A et B	-0,29	0,35	-0,51
4F : A et L	-0,16	-0,08	-0,21
4F : B et L	0,05	0,71	0,15

Bien que certaines corrélations soient très fortes, il est difficile d'accepter ces valeurs *prima facie*, car ces coefficients de corrélations sont calculés à partir de très petits échantillons, soit 10 paires de valeurs pour le délai 0 et 9 paires pour les délais -1 et +1. Les corrélations sont particulièrement étonnantes au délai 0 pour 4F, pour qui la corrélation entre B et L est deux fois plus élevée que la corrélation entre A et B, alors que les notes A et B concernent les habiletés communicationnelles. Logiquement, les niveaux de sévérité A et B devraient être plus fortement corrélés que les niveaux A et L ou B et L. Le fait que, pour les 2 examinateurs, la corrélation entre B et L soit la plus forte est inexplicable, surtout que ces corrélations sont très fortes (0,73 et 0,71). Il est difficile de conclure quoi que ce soit de ces coefficients.

4.3.4 Corrélations croisées interindividuelles

Les corrélations croisées interindividuelles permettent d'atteindre le troisième objectif spécifique de recherche, qui vise à « comparer l'évolution du niveau de sévérité d'examinateurs travaillant ensemble ». Durant cette période de 30 mois, 1H et 4F ont évalué ensemble 170 candidats, répartis sur les 10 temps de mesure, pour une moyenne de 17 candidats par temps. Ils ont évalué ensemble au moins 1 candidat pour chacun des 10 temps de mesure. Le tableau 4.18 présente les coefficients de corrélation croisée interindividuelle, pour chacune des 3 notes, entre 1H et 4F.

Tableau 4.18
Corrélations croisées interindividuelles aux délais -1, 0 et +1 pour les examinateurs 1H et 4F, du 2010-10 au 2013-03

	Délai		
	-1	0	+1
A	0,23	-0,86	0,11
B	0,02	-0,19	-0,36
L	-0,04	0,26	-0,40

Ces corrélations étant calculées à partir de 10 paires de données (délai 0) ou de 9 (délais ±1), il faut considérer les résultats avec circonspection. En particulier, la force de la corrélation au délai 0 pour A est étonnante, mais elle rend bien compte de l'évolution des courbes de la figure 4.17, où un parallélisme inversé est évident. Remarquons que les corrélations sont négatives pour les séries A et B, pour lesquelles le niveau de sévérité moyen des deux examinateurs est plus éloigné, tandis que la corrélation pour L, où les niveaux moyens de sévérité sont quasi identiques pour les premiers temps de mesure, est positive. Cela suggère que ces deux examinateurs ont, consciemment ou non, cherché à se rapprocher, surtout pour A, où leur niveau de sévérité sont le plus éloigné, la différence entre leur niveau de sévérité au temps 1 étant de 0,87 logit, une différence importante. Si les corrélations au délai -1, donc entre le niveau de 1H au temps $t - 1$ et celui de 4F au temps t, semblent nulles, les corrélations au délai +1, elles, sont assez importantes dans 2 cas sur 3. *Stricto sensu*, cela signifie que, pour B et L, le niveau de sévérité de 4F, au temps $t - 1$, est négativement corrélé au niveau de 1H au temps t. Toutefois, pour le cas présent où chaque temps de mesure équivaut à une période de 3 mois, cette interprétation n'est pas acceptable. L'idée selon laquelle un examinateur « ajusterait » ses notes (et donc son niveau de sévérité) en fonction des notes accordées par un autre examinateur 3 mois auparavant ne peut être retenue. Pour conclure, le faible nombre de temps de mesure invite à la prudence, mais il semble qu'il y ait une relation entre le niveau de

sévérité de ces deux examinateurs au cours de cette période, du moins pour les séries A.

4.3.5 Comparaisons débutants et expérimentés

L'examinatrice 4F était débutante au cours de cette période, aussi il est possible de comparer l'évolution de son niveau de sévérité à celui de 1H, l'autre examinateur étudié ici. Il y a toutefois une limite importante en ce que, pour ces données, chaque temps de mesure représente une période de 3 mois et le nombre de candidats évalués est fortement asymétrique ; les premiers temps de mesure de cette période incluent très peu d'évaluations, alors que les 3 derniers en ont énormément. Pour les 7 premiers temps, l'examinatrice 4F a évalué 130 candidats pour une moyenne de 19 candidats par temps de mesure, mais 531 candidats, pour une moyenne de 177 candidats, durant les 3 derniers temps de mesure. En retenant le même seuil qu'à la section 4.2.4, soit l'évaluation des 100 premiers candidats, l'examinatrice 4F a donc été « débutante » pour les 6 premiers temps de cette période, au cours desquels elle a évalué 97 candidats. Pour les séries B et L, la comparaison entre 1H et 4F révèle une seule différence intéressante. Pour les 6 premiers temps de ces 2 séries, 4F est plus clémente que 1H 11 temps sur 12 et l'écart de sévérité entre les deux examinateurs est faible pour B et minuscule pour L, comme le montre les figures 4.18 et 4.19. La différence réside en la série B de 4F par rapport à 1H ; la différence moyenne de valeur entre les niveaux de sévérité successifs[19], en valeur absolue, est de 0,16 logit pour 4F et 0,07 logit pour 1H. En d'autres mots, le niveau de sévérité de 4F est moins stable que celui de 1H pour les 6 premiers temps de mesure, il est un peu plus volatil.

Les séries A, elles, révèlent des contrastes intéressants entre le niveau de sévérité de 1H et 4F. C'est le seul des 3 cas où l'écart de sévérité entre les deux examinateurs est assez important. Cet écart est au maximum au temps 1, où 0,87 logit sépare les deux

[19] C'est-à-dire la différence entre t_2 et t_1, t_3 et t_2...

examinateurs. Cet écart diminue à 0,25 logit au temps 2, puis rebondit à 0,83 et 0,73 logit lors des 2 temps suivants, avant de se résorber durablement. L'autre résultat notable est que le niveau de sévérité de 4F est plus volatil que celui de 1H, la différence moyenne de valeur entre les niveaux de sévérité successifs, en valeur absolue, étant de 0,29 logit pour 4F et 0,14 logit pour 1H. Pour conclure, certains indices montrent, surtout pour A, que l'examinatrice débutante 4F a un niveau de sévérité plus instable que celui de son collègue plus expérimenté.

4.4 Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du temps chronologique, du 2011-09 au 2013-02

4.4.1 Représentation graphique et description

Les figures 4.20 à 4.22 montrent les graphiques chronologiques des séries chronologiques des examinateurs 1H et 5F, du 2010-09 au 2013-02. Chaque temps de mesure représente une période de ½ mois et il y a 34 temps de mesure. Au cours de cette période, 1H a évalué 435 candidats et 5F en a évalué 587, pour des moyennes respectives de candidats évalués par temps de mesure de 13 et 17. Les estimations de niveau de sévérité, pour chaque note, sont communes pour les 2 examinateurs et leurs niveaux de sévérité peuvent donc être directement comparés. Les figures montrent les niveaux de sévérité dans un même graphique. Afin de favoriser la comparaison des fluctuations entre les séries des 3 notes, l'étendue de l'ordonnée est la même pour les 3 graphiques, soit 1,20 logit. Les lignes pointillées représentent la tendance linéaire globale de chaque série. Précisons, puisque ce n'est pas clair dans les graphiques, que la tendance linéaire globale de 1H est au-dessus de la tendance de 5F, et ce dans les 3 graphiques.

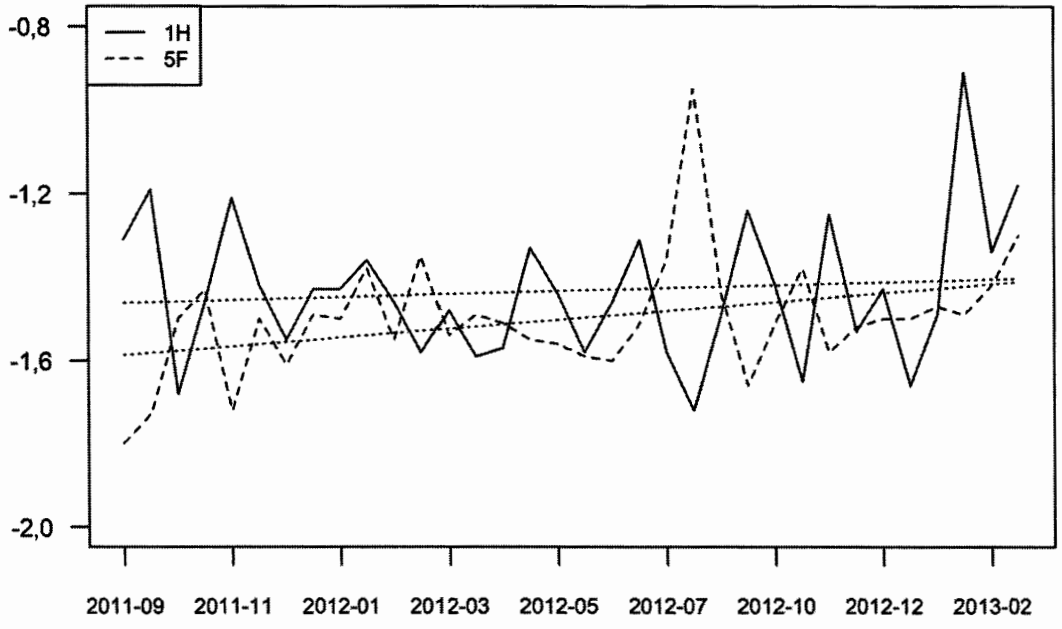


Figure 4.20 : Niveaux de sévérité A des examinateurs 1H et 5F, du 2011-09 au 2013-02. L'ordonnée est en logit.

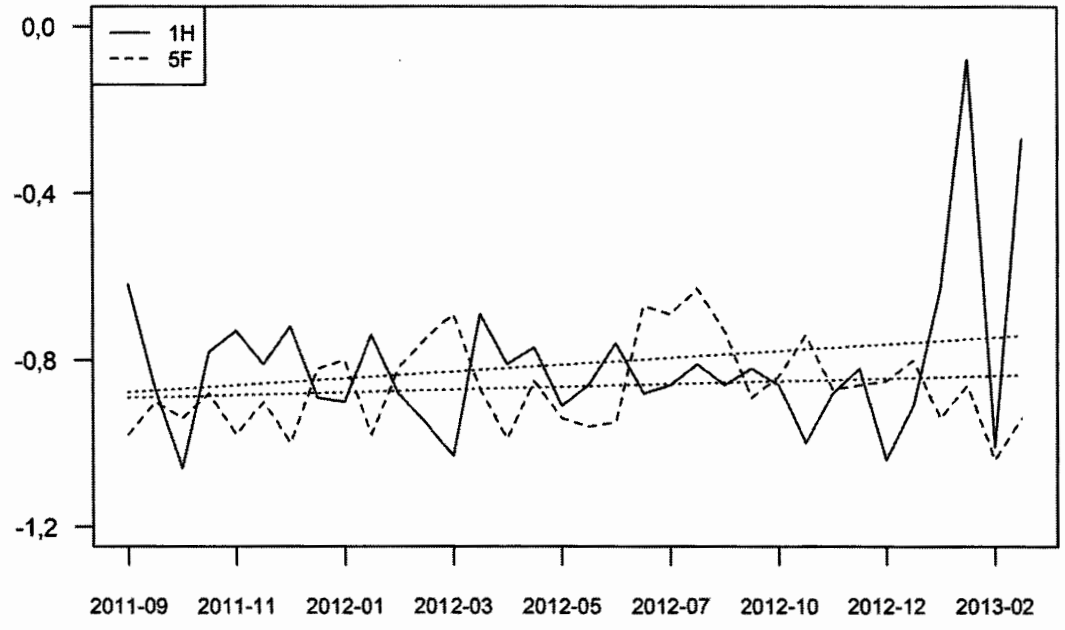


Figure 4.21 : Niveaux de sévérité B des examinateurs 1H et 5F, du 2011-09 au 2013-02. L'ordonnée est en logit.

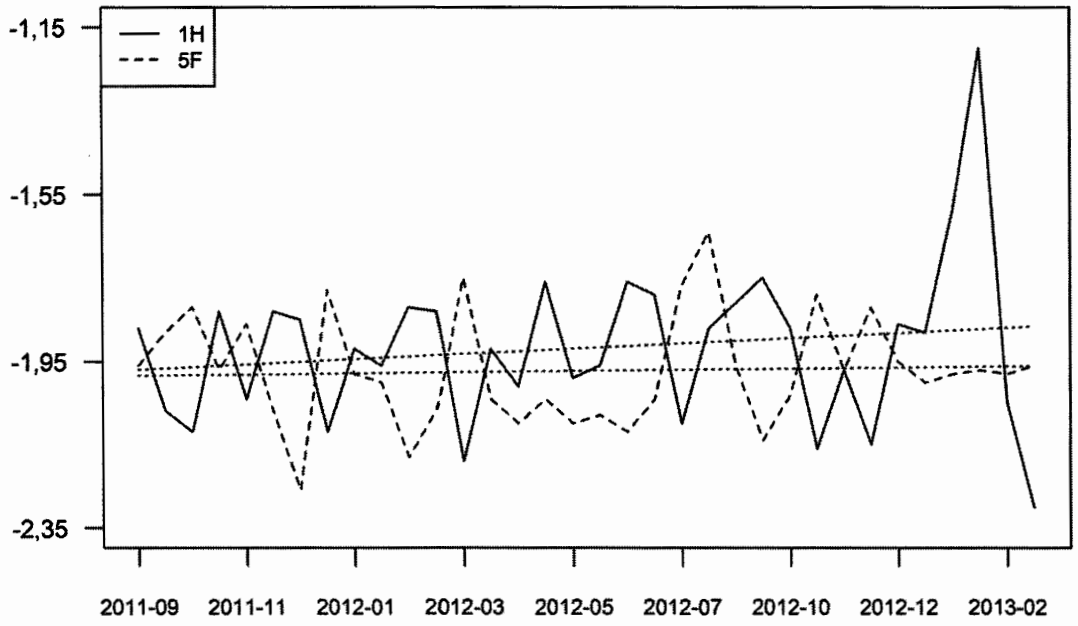


Figure 4.22 : Niveaux de sévérité L des examinateurs 1H et 5F, du 2011-09 au 2013-02. L'ordonnée est en logit.

Les valeurs extrêmes dans les 2 derniers mois des séries de 1H s'expliquent probablement par le faible nombre de candidats évalués par cet examinateur en janvier (6) et février 2013 (10). Puisque chaque temps de mesure représente ½ mois, les niveaux de sévérité de ces 4 derniers temps ont donc été estimés à partir de 3 à 5 candidats évalués, ce qui est suffisamment faible pour expliquer ces valeurs extrêmes. La présence de ces valeurs extrêmes explique l'important plateau de la série B, d'une durée de 6 mois, de juin à décembre, où le niveau de sévérité de 1H est sous sa tendance linéaire globale. Il y a aussi quelques plateaux d'une durée d'environ 3 mois, par exemple pour la série A de 5F, de novembre 2012 à février 2013, ou pour la série L de 5F, de mars à juin 2012. Dans tous ces cas, l'écart entre les plateaux et la tendance linéaire globale est assez faible, environ 0,20 logit. La valeur extrême de la série A de 5F est, elle, difficile à expliquer. Cette valeur correspond à la deuxième moitié de juillet 2012, période durant laquelle 5F a évalué 13 candidats, ce qui est un nombre suffisamment élevé pour que la valeur élevée du niveau de sévérité ne soit

pas causée par une ou deux évaluations inhabituelles. Cette valeur reflète donc une augmentation locale du niveau de sévérité de 5F. Pour ces 6 séries, les valeurs en scores standardisés (soit la valeur, en écarts types, de l'éloignement de la moyenne) les plus grandes, en valeur absolue, sont systématiquement des valeurs positives, montrant une sévérité plus importante. Considérant la période complète de 18 mois, les niveaux de sévérité des 2 examinateurs sont stables, les tendances linéaires globales ayant une très faible pente. La tendance linéaire globale de la série 5F A a la pente la plus importante, mais la différence entre ses valeurs finales et initiales est de 0,18 logit, ce qui est peu. Les niveaux de sévérité des deux examinateurs s'entrecroisent, sans qu'un patron clair n'émerge. Ainsi, 1H est plus sévère que 5F 22 fois sur 34 pour A, mais seulement 15 fois sur 34 pour B et 21 fois sur 34 pour L. De manière générale, les 6 séries évoluent en dents de scie, avec de constants soubresauts au-dessus et en dessous de la tendance linéaires globales, les tendances linéaires locales étant très rares. Outre les plateaux susmentionnés, notons la tendance linéaire locale descendante de la série 5F A, de mars à juin 2012, et la tendance linéaire locale ascendante de la même série, de novembre 2011 à février 2012.

Lorsque les données de chaque série sont considérées conjointement, on constate que les séries L et la série 5F B sont approximativement normalement distribuées, mais que les écarts à la normalité sont plus importants pour les 2 séries A et la série 1H B, les valeurs extrêmes précédemment mentionnées gonflant l'asymétrie positive. Les diagrammes quantile-quantile de la figure 4.23 illustrent l'adéquation à la distribution normale. Le tableau 4.19, lui, contient les statistiques descriptives des données des 6 séries.

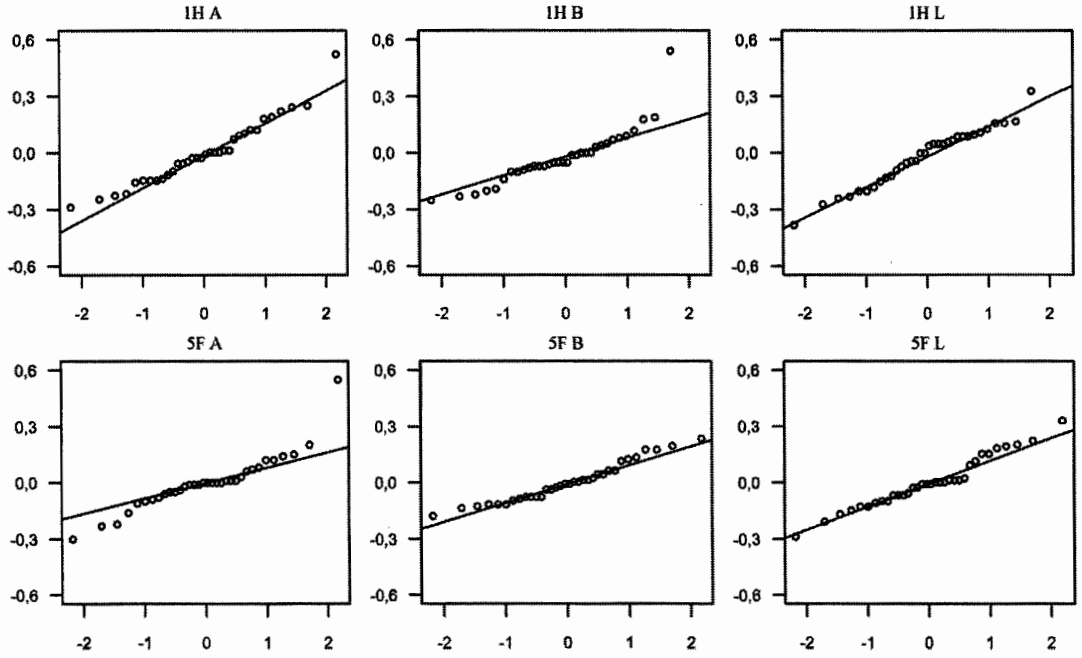


Figure 4.23 : Diagrammes quantile-quantile des séries de 1H et 5F, du 2011-09 au 2013-02

Tableau 4.19
Statistiques descriptives des séries chronologiques de 1H et 5F, du 2011-09 au 2013-02

Série	*m*	*s*	min	max	étendue	asymétrie	aplatissement	% ±2 *s*
1H A	-1,43	0,17	-1,72	-0,91	0,81	0,75	0,72	97
5F A	-1,50	0,15	-1,80	-0,95	0,85	1,22	4,23	94
1H B	-0,81	0,19	-1,06	-0,08	0,98	1,98	4,79	94
5F B	-0,86	0,10	-1,04	-0,63	0,41	0,47	-0,71	97
1H L	-1,92	0,20	-2,30	-1,20	1,10	1,09	3,03	97
5F L	-1,97	0,13	-2,26	-1,64	0,62	0,35	-0,22	94

% ±2 *s* : pourcentage des valeurs à plus ou moins 2 écarts types de la moyenne

Relevons le pourcentage important de valeurs comprises à ±2 écarts types de la moyenne, ce qui montre bien que les données sont regroupées autour de la moyenne et que les valeurs extrêmes sont rares. Ces pourcentages sont aussi très près du

pourcentage de données à ±2 écarts types d'une distribution parfaitement normale, soit 95,45 %.

4.4.2 Modélisation AMMI

Sur 6 séries, 3 ne sont que du bruit blanc, soit les séries 1H A, 1H B et 5F L. Les modèles AMMI décrivant le mieux les 3 autres séries sont présentés dans le tableau 4.20.

Tableau 4.20
Modèles AMMI des séries chronologiques de 1H et 5F, du 2011-09 au 2013-02

Série	AMMI	Coefficients*	Erreurs types
1H L	0,0,2	0,04 ; -0,46	0,18 ; 0,19
5F A	0,0,1	0,30	0,15
5F B	1,0,0	0,39	0,16

*Les coefficients suivent l'ordre $A_1...A_p$, $MM_1...MM_q$ et sont séparés par un « ; »

Dans les trois cas, les conditions d'utilisation de la modélisation AMMI sont respectées (voir les résultats des tests diagnostiques en annexe A). À l'instar des modèles obtenus avec les séries chronologiques établies selon le nombre de candidats évalués, l'ordre des paramètres p et q est petit, soit 1 ou 2, et les séries ont un paramètre p ou q, mais pas les deux. Les coefficients obtenus ici sont d'ampleur moyenne et les deux modèles de l'examinatrice 5F ont des coefficients d'ampleur similaire. Remarquons que les séries de ces 2 examinateurs ne correspondent pas vraiment aux modèles obtenus à la section 4.2, lorsque l'évolution du niveau de sévérité de ces examinateurs est modélisée en fonction du nombre de candidats qu'ils ont évalués. Les séries 1H L et 5F B sont, selon le découpage du temps de la section 4.2, représentées par le modèle nul, alors qu'elles sont ici représentées par un modèle 0,0,2 ou 1,0,0. C'est intéressant, car cela fait ressortir l'ambiguïté inhérente à la démarche de modélisation de l'évolution temporelle du niveau de sévérité. Puisqu'il

n'y a pas d'unité « naturelle » de base pour conceptualiser le temps – par exemple, pour les données économiques, le jour ouvrable – les mêmes données peuvent, selon un découpage temporel différent, mener à des modèles différents. Finalement, les 3 séries ayant un modèle non nul sont également réparties entre les données A, B et L.

4.4.3 Corrélations croisées intraindividuelles

Le tableau 4.21 énumère les coefficients de corrélations croisées intraindividuelles des examinateurs 1H et 5F aux délais -1, 0 et +1.

Tableau 4.21
Corrélations croisées intraindividuelles aux délais -1, 0 et +1 pour les examinateurs 1H et 5F, du 2011-09 au 2013-02

	Délai		
	-1	0	+1
1H : A et B	-0,15	0,59	0,27
1H : A et L	-0,23	0,42	0,30
1H : B et L	0,11	0,54	0,13
5F : A et B	0,10	0,39	0,13
5F : A et L	0,09	0,40	0,08
5F : B et L	0,27	0,40	-0,12

En accord avec les résultats observés à la section 4.2, les coefficients aux délais -1 et +1 ne révèlent aucun patron systématique et ils sont plus ou moins distribués autour de 0. Cela montre que le niveau de sévérité d'une note au temps t ne permet pas de prédire le niveau de sévérité des deux autres notes au temps $t + 1$. En revanche, les niveaux de sévérité des trois notes au temps t sont positivement corrélés, et les coefficients sont remarquablement stables : les 3 coefficients de 1H ont une étendue de 0,17 et les 3 coefficients de 5F sont égaux à 0,01 près. Il est encore une fois étrange de constater que les 3 coefficients de chaque examinateur sont à peu près égaux et que les coefficients entre A et B ne sont pas plus grand que les 2 autres

coefficients, ce qui va à l'encontre de la logique attendue étant donné la similarité conceptuelle entre A et B.

4.4.4 Corrélations croisées interindividuelles

Durant cette période de 18 mois, 1H et 5F ont évalué ensemble 284 candidats, répartis sur les 34 temps de mesure, pour une moyenne de 8 candidats par temps. À noter que ces deux examinateurs n'ont évalué aucun candidat en commun lors des 4 derniers temps de mesure, mais ils ont évalué ensemble au moins 1 candidat lors des 30 premiers temps de mesure. Le tableau 4.22 présente les coefficients de corrélation croisée interindividuelle, pour chacune des 3 notes, entre 1H et 5F.

Tableau 4.22
Corrélations croisées interindividuelles aux délais -1, 0 et +1 pour les examinateurs 1H et 5F, du 2011-09 au 2013-02

	Délai		
	-1	0	+1
A	0,09	-0,39	-0,07
B	-0,19	-0,17	-0,33
L	-0,05	-0,31	0,08

Les corrélations au délai 0 sont toutes négatives et d'ampleur faible ou moyenne, ce qui indique une tendance générale au rapprochement du niveau de sévérité des deux examinateurs, puisqu'une hausse de l'un est accompagnée d'une baisse de l'autre. Ces corrélations négatives participent au phénomène du chevauchement et des croisements des deux courbes, visibles sur les figures 4.20 à 4.22, puisque les niveaux moyens de sévérité des deux examinateurs sont très près (0,05 à 0,07 logit de différence). Ces corrélations ont toutefois le défaut d'effacer certains patrons locaux très intéressants, patrons qui sont aisément remarqués lors d'un examen visuel des graphiques chronologiques. Par exemple, si l'on ne considère, pour les séries B, que les temps 4 à 14, la corrélation croisée entre les niveaux de sévérité au délai 0 est de -

0,90. De même, si l'on ne retient que les temps 1 à 14 des séries L, la corrélation au délai 0 est de -0,87. Un autre exemple de ces dynamiques est l'évolution du niveau de sévérité A, entre les temps 6 et 11. Comme il n'y a que 6 temps de mesure, calculer une corrélation n'aurait pas de sens, mais l'examen visuel révèle une symétrie quasi parfaite entre les niveaux de sévérité des deux examinateurs. Bien sûr, ces exemples sont sciemment choisis afin d'obtenir des coefficients très forts, mais il y a clairement des dynamiques locales importantes entre les niveaux de sévérité de ces deux examinateurs. La corrélation croisée des temps 1 à 14 des séries L représente tout de même 7 mois de travail commun, au cours desquels 1H et 5F ont conjointement évalué 122 candidats, soit ~9 candidats évalués par période de ~15 jours. Les corrélations aux délais ±1 semblent nulles, à l'exception de la corrélation B au délai +1, un peu plus importante. Il est intéressant de constater que les corrélations aux délais -1 et 0 sont plus faibles, mais il est difficile de supposer un lien véritable entre le niveau de sévérité de 5F au temps t et celui de 1H au temps $t + 1$, mais seulement pour les notes B et pas pour les deux autres notes.

4.5 Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du temps chronologique, du 2012-06 au 2013-11

4.5.1 Représentation graphique et description

Les figures 4.24 à 4.26 montrent les graphiques chronologiques des séries chronologiques des examinateurs 4F et 5F, du 2012-06 au 2013-11. Chaque temps de mesure représente une période de ½ mois et il y a 36 temps de mesure. Au cours de cette période, 4F a évalué 1051 candidats et 5F en a évalué 973, pour des moyennes respectives de candidats évalués par temps de mesure de 29 et 27. Les estimations de niveau de sévérité, pour chaque note, sont communes pour les 2 examinateurs et leurs niveaux de sévérité peuvent donc être directement comparés. Les figures montrent les niveaux de sévérité dans un même graphique. Afin de favoriser la comparaison des

fluctuations entre les séries des 3 notes, l'étendue de l'ordonnée est la même pour les 3 graphiques, soit 1,25 logit.

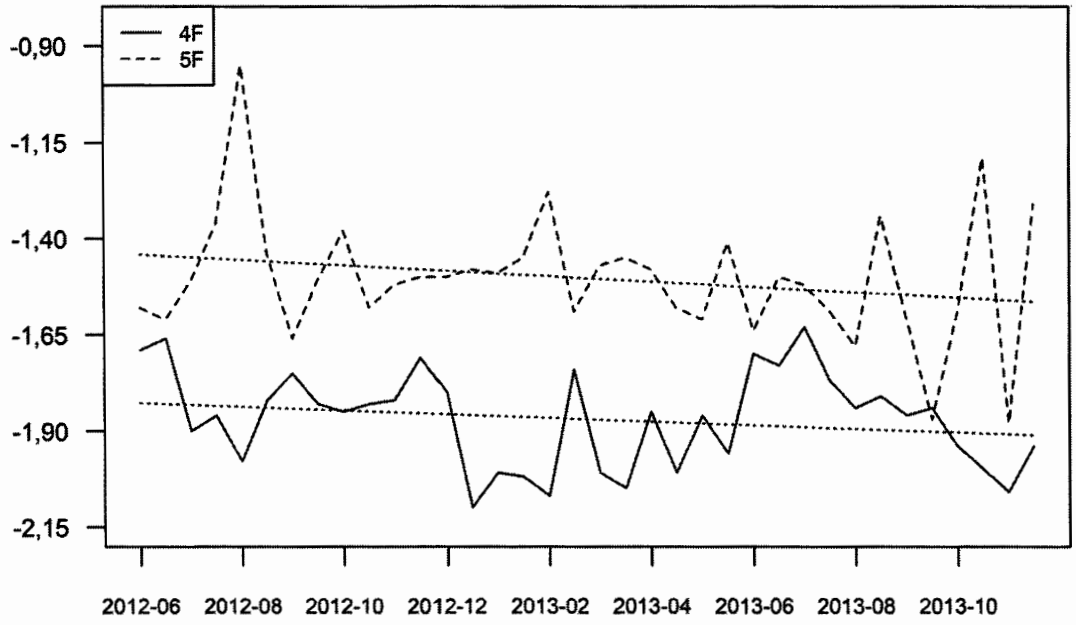


Figure 4.24 : Niveaux de sévérité A des examinatrices 4F et 5F, du 2012-06 au 2013-11. L'ordonnée est en logit.

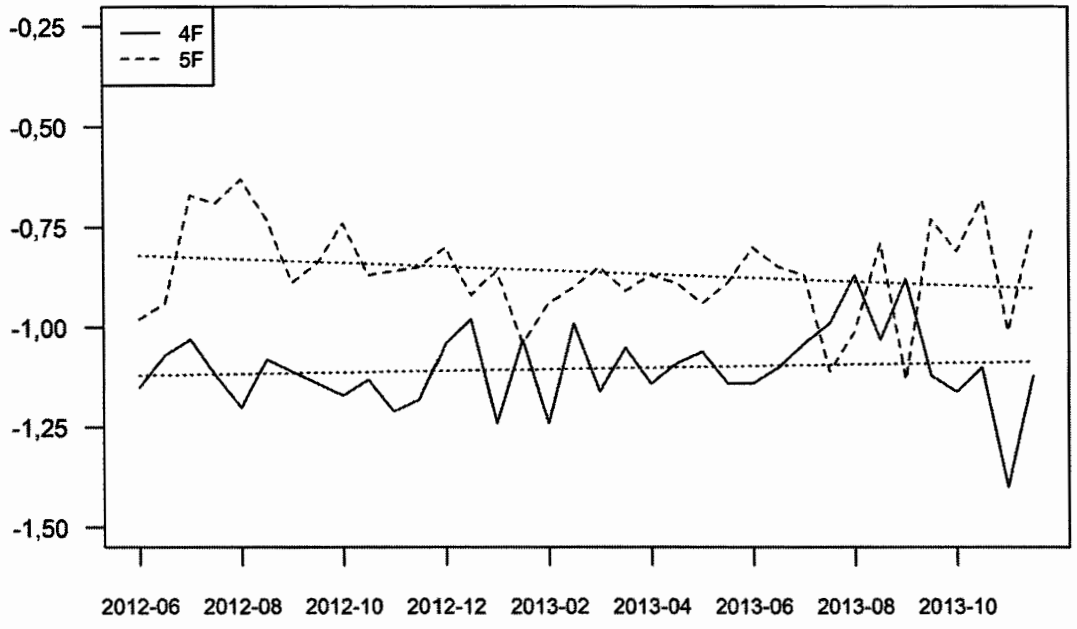


Figure 4.25 : Niveaux de sévérité B des examinatrices 4F et 5F, du 2012-06 au 2013-11. L'ordonnée est en logit.

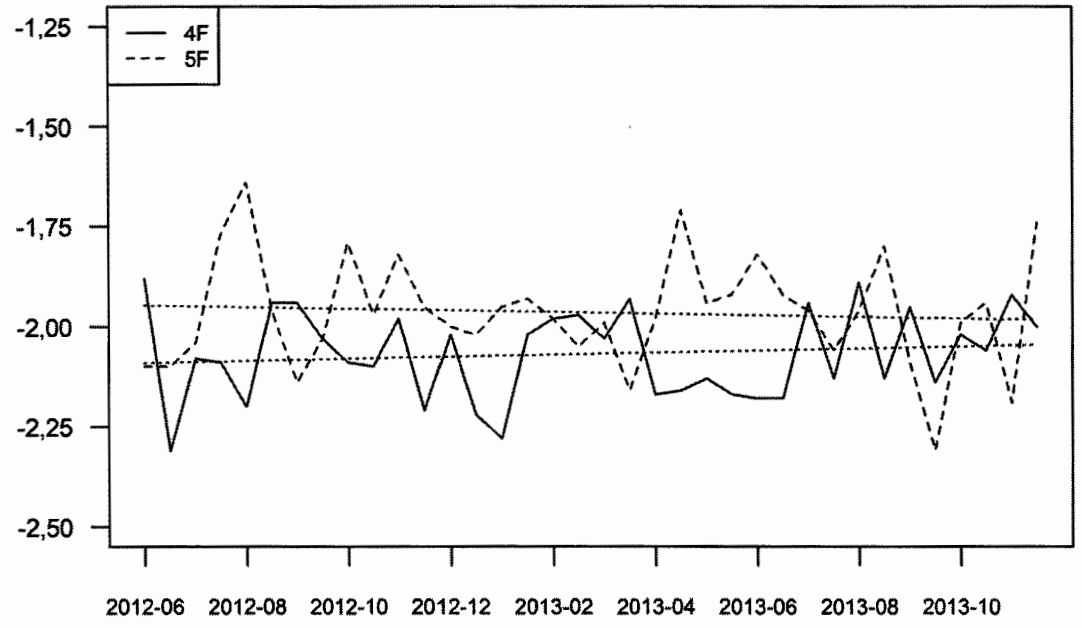


Figure 4.26 : Niveaux de sévérité L des examinatrices 4F et 5F, du 2012-06 au 2013-11. L'ordonnée est en logit.

La valeur extrême de la série A de l'examinatrice 5F, en août 2012, résulte du fait que cette examinatrice a seulement évalué 4 candidats au cours de ce mois. Les maxima des séries B et L de cette examinatrice sont aussi durant ce mois. Concernant la période complète de 18 mois, les deux examinatrices ont un niveau de sévérité stable. La pente la plus importante parmi les tendances linéaires globales des 6 séries, soit la pente de la série 5F A, ne représente qu'une différence de -0,12 logit entre le premier et le dernier temps de la série. L'ordre des niveaux de sévérité entre les 2 examinatrices est constant, 5F étant plus sévère 35 fois sur 36 pour les scores A, 32 fois pour B et 24 fois pour L, où l'ordre permute davantage. Notons la présence d'un plateau dans la série A de 4F, de juin à septembre 2013, où le niveau de sévérité est supérieur à la tendance linéaire globale. Une configuration en dents de scie est présente sur certaines périodes, dans chacune des 6 séries. Il y a également 6 tendances linéaires locales assez lisses réparties ainsi parmi les 6 séries : de novembre 2012 à février 2013 pour la série 5F A, de juin à novembre 2013 pour 4F A, de juin à novembre 2012 et de mai à août 2013 pour 4F B, de février à juin 2013 pour 5F B et de novembre 2012 à mars 2013 pour 5F L.

Lorsque les données sont considérées de manière synchrone, sans ordre temporel, on constate que 5 séries sur 6 suivent approximativement une distribution normale, seule la série 5F A étant franchement anormale. Il y a tout de même 3 séries ayant un coefficient d'aplatissement assez élevé (4F B) ou bas (4F A et L), mais ces séries sont symétriques et les excès d'aplatissement ne sont pas trop importants. Les écarts les plus importants à la moyenne, en score standardisé, sont positifs, pour les 2 séries A, et négatifs pour les séries B et L. En d'autres termes, ces examinatrices ont parfois un excès de sévérité pour les scores A et un excès de clémence pour les scores B et L. Les pourcentages de données situées à ±2 écarts types de la moyenne sont assez près du pourcentage d'une distribution normale, soit 95,45 %, à l'exception de la série 4F A, où toutes les données sont à moins de 2 écarts types de la moyenne. La figure 4.27

illustre les diagrammes quantile-quantile des 6 séries et le tableau 4.23, lui, contient les statistiques descriptives de ces séries.

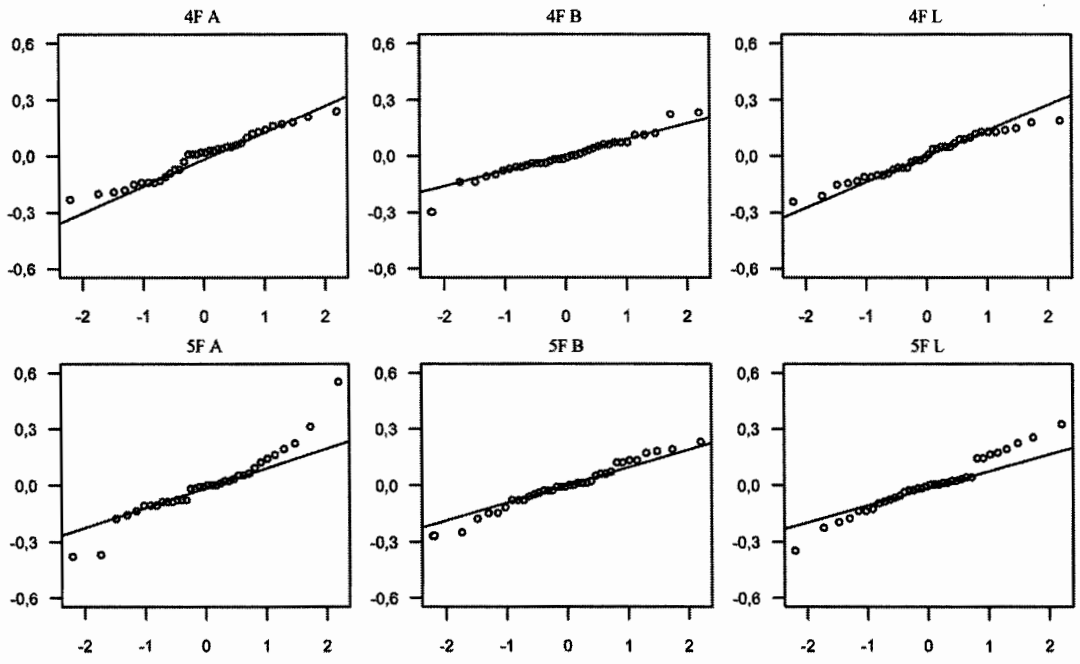


Figure 4.27 : Diagrammes quantile-quantile des séries de 4F et 5F, du 2012-06 au 2013-11

Tableau 4.23
Statistiques descriptives des séries chronologiques de 4F et 5F, du 2012-06 au 2013-11

Série	*m*	*s*	min	max	étendue	asymétrie	aplatissement	% ±2 *s*
4F A	-1,87	0,13	-2,10	-1,63	0,47	-0,05	-1,07	100
5F A	-1,50	0,17	-1,88	-0,95	0,93	0,61	2,08	92
4F B	-1,10	0,10	-1,40	-0,87	0,53	-0,13	1,19	92
5F B	-0,86	0,12	-1,13	-0,63	0,50	-0,17	-0,31	94
4F L	-2,07	0,11	-2,31	-1,88	0,43	-0,18	-1,02	97
5F L	-1,96	0,14	-2,31	-1,64	0,67	0,10	0,02	94

% ±2 *s* : pourcentage des valeurs à plus ou moins 2 écarts types de la moyenne

4.5.2 Modélisation AMMI

Sur les 6 séries, 3 sont impossibles à distinguer du bruit blanc. Les modèles AMMI décrivant le mieux les 3 autres séries, 4F A, 4F B et 5F L, sont présentés dans le tableau 4.24.

Tableau 4.24
Modèles AMMI, des séries chronologiques de 4F et 5F, du 2012-06 au 2013-11

Série	AMMI	Coefficients*	Erreurs types
4F A	1,0,0	0,40	0,16
4F B	0,0,2	0,02 ; 0,44	0,16 ; 0,17
5F L	1,0,2	0,53 ; -0,34 ; -0,61	0,17 ; 0,16 ; 0,15

*Les coefficients suivent l'ordre A_1...Ap, MM_1...MMq et sont séparés par un « ; »

Les modèles respectent tous les conditions d'utilisation de la modélisation AMMI (voir les résultats des tests diagnostiques en annexe A). Encore une fois, l'ordre est petit, soit 1 ou 2, et les coefficients de l'ordre le plus élevé sont de taille moyenne, allant de 0,40 à 0,61 en valeur absolue. Notons la présence du modèle 1,0,2, de la série 5F L, où il y a à la fois un coefficient autorégressif et deux coefficients de moyenne mobile, ce qui est plus rare. Comme pour l'ensemble de données précédent, de la section 4.4.2, les 3 séries ayant un modèle non nul se répartissent également en A, B et L.

4.5.3 Corrélations croisées intraindividuelles

Le tableau 4.25 contient les coefficients de corrélation croisés intraindividuelles des examinateurs 4F et 5F aux délais -1, 0 et +1.

Tableau 4.25
Corrélations croisées intraindividuelles aux délais -1, 0 et +1 pour les examinateurs 4F et 5F, du 2012-06 au 2013-11

	Délai		
	-1	0	+1
4F : A et B	0,30	0,14	-0,16
4F : A et L	-0,11	0,04	0,13
4F : B et L	-0,17	0,17	-0,05
5F : A et B	-0,15	0,49	0,16
5F : A et L	-0,25	0,54	0,13
5F : B et L	-0,05	0,33	-0,18

La faiblesse des coefficients des 3 séries de l'examinatrice 4F au délai 0 est très étonnante. Alors que les valeurs des 3 coefficients de 5F ressemblent beaucoup aux valeurs observées des coefficients avec les ensembles de données étudiés ci-dessus, les 3 coefficients de 4F sont inexplicables, surtout pour la corrélation entre A et B, où les deux notes représentent l'évaluation d'habiletés communicationnelles. Il est donc logique de supposer que les niveaux de sévérité de ces deux notes soient positivement corrélés, ce qui n'est pas le cas ici. Ces 6 séries ayant 36 temps de mesure, on ne peut expliquer la faiblesse de ces coefficients par le faible nombre de paires de données. Les 3 coefficients de l'examinatrice 5F sont homogènes, mais relevons l'étonnante force du coefficient entre A et L, légèrement supérieur au coefficient entre A et B, ce qui contrevient à la logique attendue. Pour terminer, remarquons que, comme le montrent les valeurs des corrélations aux délais -1 et +1, il n'est pas possible de bien prédire les valeurs d'une série d'une examinatrice au temps t à l'aide des valeurs d'une des deux autres séries de la même examinatrice au temps $t-1$.

4.5.4 Corrélations croisées interindividuelles

Durant cette période de 18 mois, 4F et 5F ont évalué ensemble 355 candidats, répartis sur les 36 temps de mesure, pour une moyenne de ~10 candidats par temps. À noter

que ces deux examinatrices n'ont évalué aucun candidat en commun lors des 5 derniers temps de mesure, mais elles ont évalué ensemble au moins 1 candidat lors des 30 premiers temps de mesure. Le tableau 4.26 présente les coefficients de corrélation croisée interindividuelle, pour chacune des 3 notes, entre 4F et 5F.

Tableau 4.26
Corrélations croisées interindividuelles aux délais -1, 0 et +1 pour les examinateurs 4F et 5F

	Délai		
	-1	0	+1
A	-0,11	-0,26	0,13
B	-0,14	-0,26	-0,22
L	-0,14	-0,22	-0,07

Les corrélations globales au délai 0 sont assez faibles et elles sont toutes trois négatives, ce qui indique une certaine tendance aux rapprochements/éloignements en dents de scie observables dans les figures 4.24 à 4.26. Si cette tendance globale est assez faible (r = -0,22 et -0,26), il y a toutefois deux dynamiques locales beaucoup plus fortes pour les séries A. La corrélation croisée entre les niveaux de sévérité 4F et 5F, des temps 1 à 10, est de -0,78, et celle des temps 11 à 20 est de -0,63. Si nous considérons les 20 premiers temps de mesure ensemble, la corrélation est tout de même de -0,50, une corrélation importante dans les circonstances, présente sur une période de 10 mois. Il semble toutefois que cette dynamique ne soit pas liée à la fréquence de travail commun de ces deux examinatrices, 4F et 5F ayant conjointement évalué 67 candidats lors des 10 premiers temps de mesure et 121 lors des temps 11 à 20, pour un total de 188 lors des temps 1 à 20. Cette somme représente 53 % des candidats conjointement évalués pour l'ensemble des 36 temps de mesure, tandis que les 20 premiers temps de mesure représentent 56 % des 36 temps de mesure de la période totale. En d'autres termes, la force des corrélations observées lors des 20 premiers temps de mesure ne s'explique pas par un nombre

proportionnellement plus élevé de candidats évalués en commun par ces deux examinatrices. Pour terminer, relevons la présence d'une corrélation négative presque aussi forte, pour les séries B, au délai +1 qu'au délai 0. Encore une fois, l'ampleur de cette corrélation est probablement le fruit du hasard, en ce que les corrélations au délai +1 des séries A et L sont près de 0, et qu'il est difficile de supposer un lien entre les notes assignées des deux examinatrices à 2 ou 3 semaines d'écart, mais seulement pour B.

4.6 Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du temps chronologique, du 2012-12 au 2013-09

4.6.1 Représentation graphique et description

Les figures 4.28 à 4.30 montrent les graphiques chronologiques des séries chronologiques des examinateurs 4F, 5F, 7F, 13H et 15F, du 2012-12 au 2013-09. Chaque temps de mesure représente ½ mois et il y a 19 temps en tout. Durant cette période, 4F a évalué 633 candidats, 5F 721, 7F 261, 13H 466 et 15F 200, pour des moyennes respectives de candidats évalués par temps de mesure de 33, 38, 14, 25 et 11. Les estimations de niveau de sévérité, pour chaque note, sont communes pour les 5 examinateurs et leurs niveaux de sévérité peuvent donc être directement comparés. Toutefois, le grand nombre d'examinateurs ne permet pas de représenter leur niveau de sévérité dans un même graphique. Chaque figure montre 5 graphiques superposés, où chaque graphique est consacré à la série d'un examinateur. Toutes les ordonnées d'une même figure ont la même échelle, et les ordonnées des 3 figures ont la même étendue, soit 1,20 logit, ce qui permet la comparaison des fluctuations pour tous les graphiques de ces 3 figures. Il faut être prudent dans l'examen visuel de ces graphiques, car les dimensions de chaque graphique sont inférieures aux dimensions des graphiques des figures précédentes, par exemple 4.24 à 4.26. Comparées aux séries de ces dernières figures, les séries qui suivent semblent plus stables et moins

dentelées, mais c'est une impression créée par la compression de l'espace graphique et par la moindre quantité de temps de mesure.

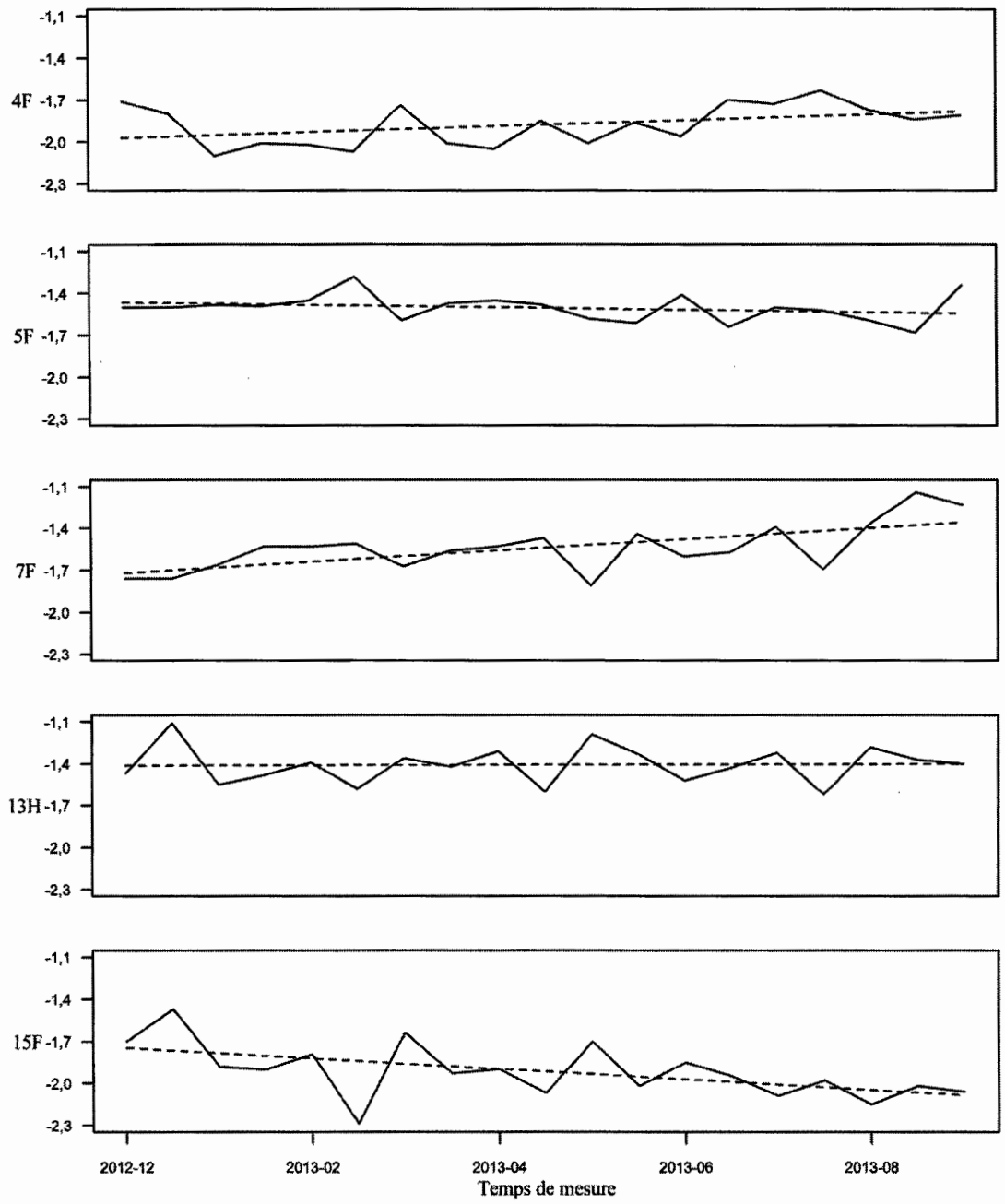


Figure 4.28 : Niveaux de sévérité A des examinateurs 4F, 5F, 7F, 13H et 15F, du 2012-12 au 2013-09. L'ordonnée est en logit.

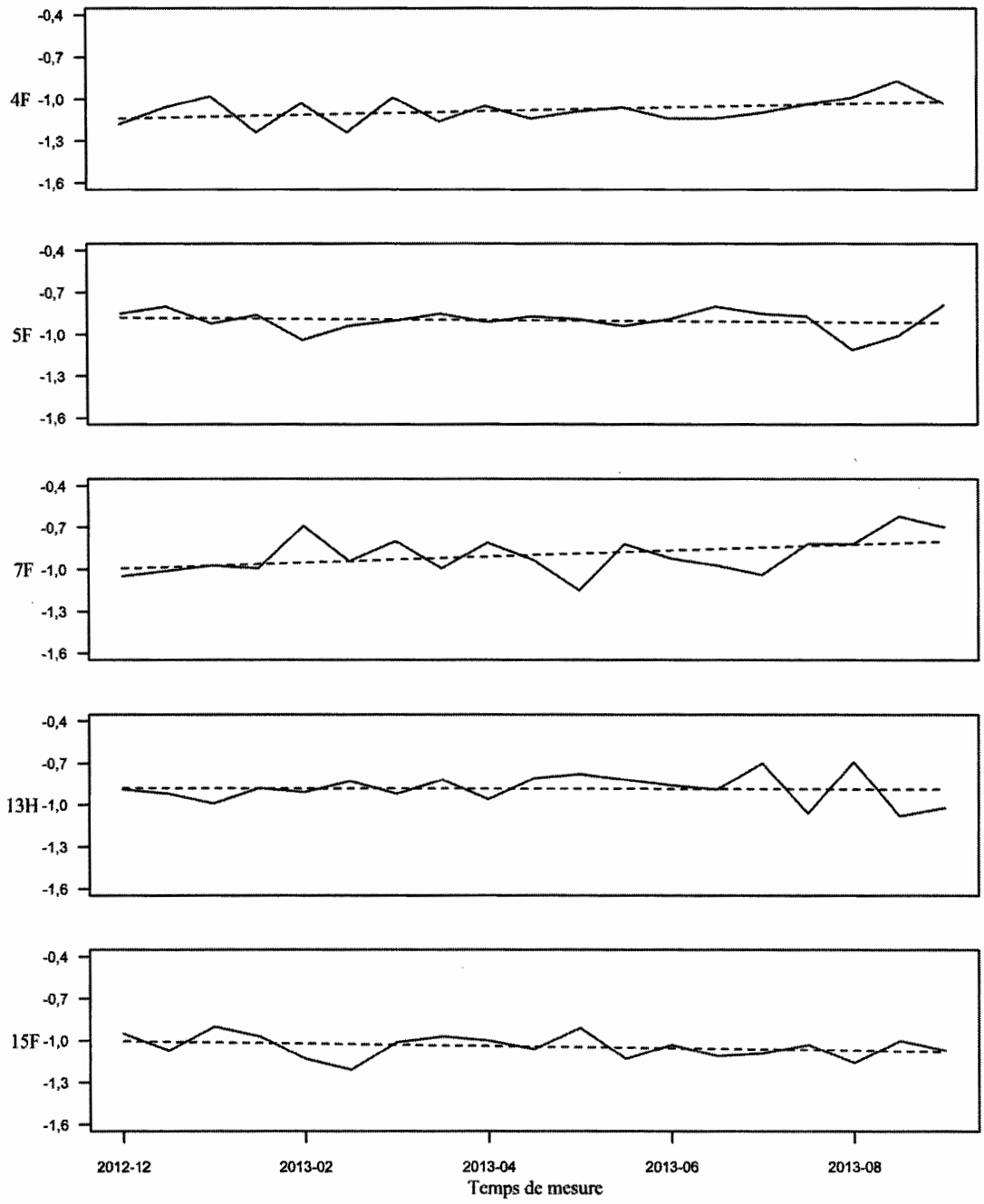


Figure 4.29 : Niveaux de sévérité B des examinateurs 4F, 5F, 7F, 13H et 15F, du 2012-12 au 2013-09. L'ordonnée est en logit.

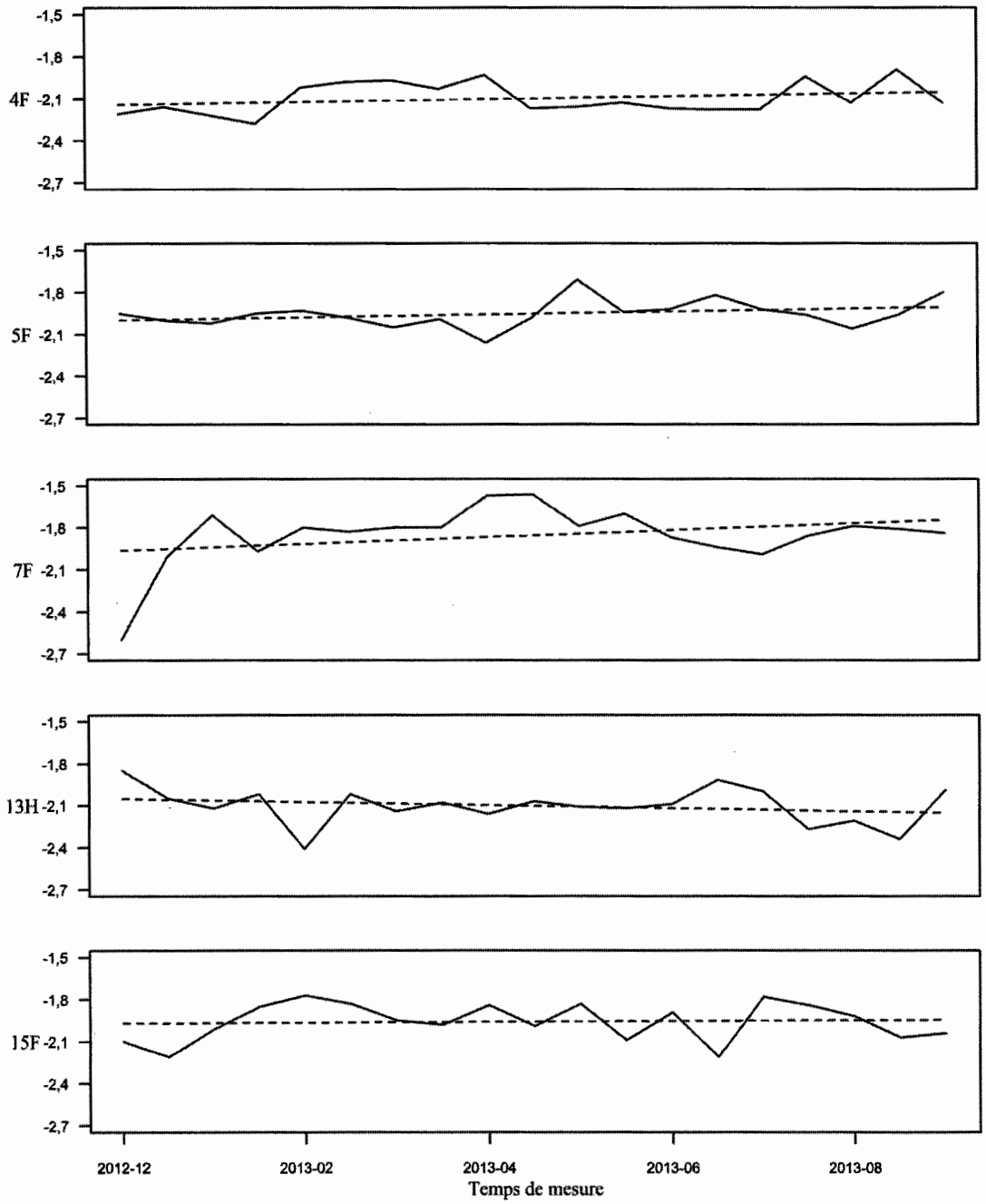


Figure 4.30 : Niveaux de sévérité L des examinateurs 4F, 5F, 7F, 13H et 15F, du 2012-12 au 2013-09. L'ordonnée est en logit.

La valeur extrême au temps 1 de la série L de l'examinatrice 7F résulte de l'évaluation de 13 candidats par celle-ci, et ne peut donc pas être expliquée par un faible nombre de données. Cette série est d'ailleurs remarquable à plusieurs titres. Le niveau de sévérité des 3 premiers temps de mesure (1,5 mois) suit une tendance linéaire locale ascendante très forte, le niveau de sévérité passant de -2,60 logits à -1,71 logit, ce qui semble indiquer une période d'ajustement au cours de laquelle le niveau de l'examinatrice 7F rejoint celui de ses 4 collègues, qui oscille entre -1,80 et -2,20 logits pour ces 3 premiers temps. S'ensuivent, pour le niveau de 7F, deux périodes de stabilité remarquable, deux plateaux d'une durée de 4 mois chacun, où le niveau de sévérité se maintient au-dessus de la tendance linéaire globale, de février à mai, pour ensuite descendre sous celle-ci et y rester jusqu'à septembre. L'autre série remarquable de cet ensemble est la série A de l'examinatrice 15F. Non seulement cette série a la tendance linéaire globale la plus forte parmi les 15 séries (5 examinateurs × 3 notes), mais les écarts à la tendance linéaire globale vont en décroissant de manière régulière tout au long des 10 mois, le niveau de sévérité des 6 derniers temps de mesure étant très près de la tendance linéaire globale. Parmi les 15 séries de cette période, une autre série possède des plateaux d'une durée minimale de 3 mois, soit la série L de l'examinatrice 4F. Le niveau de sévérité de celle-ci se maintient au-dessus de sa tendance linéaire globale de février à avril. Il passe sous cette tendance et y reste d'avril à juillet. Aucune autre série n'a de plateau ou de tendance linéaire locale forte.

En revanche, certaines séries ont une tendance linéaire globale non négligeable, 4 séries ayant une différence entre la valeur de la tendance linéaire globale au premier et au dernier temps de mesure égale ou supérieure à 0,20 logit. Il s'agit des séries 4F A, 7F A, 15F A et 7F L. Les 11 autres séries ont une tendance linéaire globale stable. Il semble que, pour ces 15 séries, il y a un lien entre le nombre de candidats évalués

et la variance des séries. Les examinatrices ayant évalué le moins de candidats (7F et 15F) sont aussi celles dont les séries ont la plus grande variance. Pour finir, l'ordre de sévérité entre les 5 examinateurs est relativement stable. Pour chacun des 3 ensembles de données, l'examinateur le plus sévère en moyenne est plus sévère que ses 4 collègues lors de 78 % des temps de mesure, le deuxième examinateur le plus sévère en moyenne est plus sévère que ses 3 collègues lors de 77 % des temps de mesure, le troisième examinateur le plus sévère l'est lors de 83 % des temps de mesure et le quatrième plus sévère est plus sévère que son collègue le moins sévère 51 % du temps. Ces pourcentages représentent la moyenne des pourcentages propres à chacun des 3 ensembles de données. C'est donc dire qu'il y a relativement peu de croisements entre le niveau de sévérité des différents examinateurs étudiés ici, sauf pour les deux examinateurs les moins sévères, dont les niveaux de sévérité permutent essentiellement au hasard.

Lorsque les données de chaque série sont considérées de manière atemporelle, la plupart sont approximativement normalement distribuées, avec quelques exceptions franches, telles les séries 4F A, 5F L, 7F L et 15F L. La figure 4.31 montre les diagrammes quantile-quantile des 15 séries, et le tableau 4.27 comprend les statistiques descriptives de ces séries. Il faut être prudent avec ces graphiques et ces statistiques, car ces 15 séries chronologiques n'ont que 19 temps de mesure, mais remarquons tout de même que les données tendent à être groupées autour de la moyenne, avec 14 séries sur 15 ayant au moins 95 % des données à moins de deux écarts types de la moyenne.

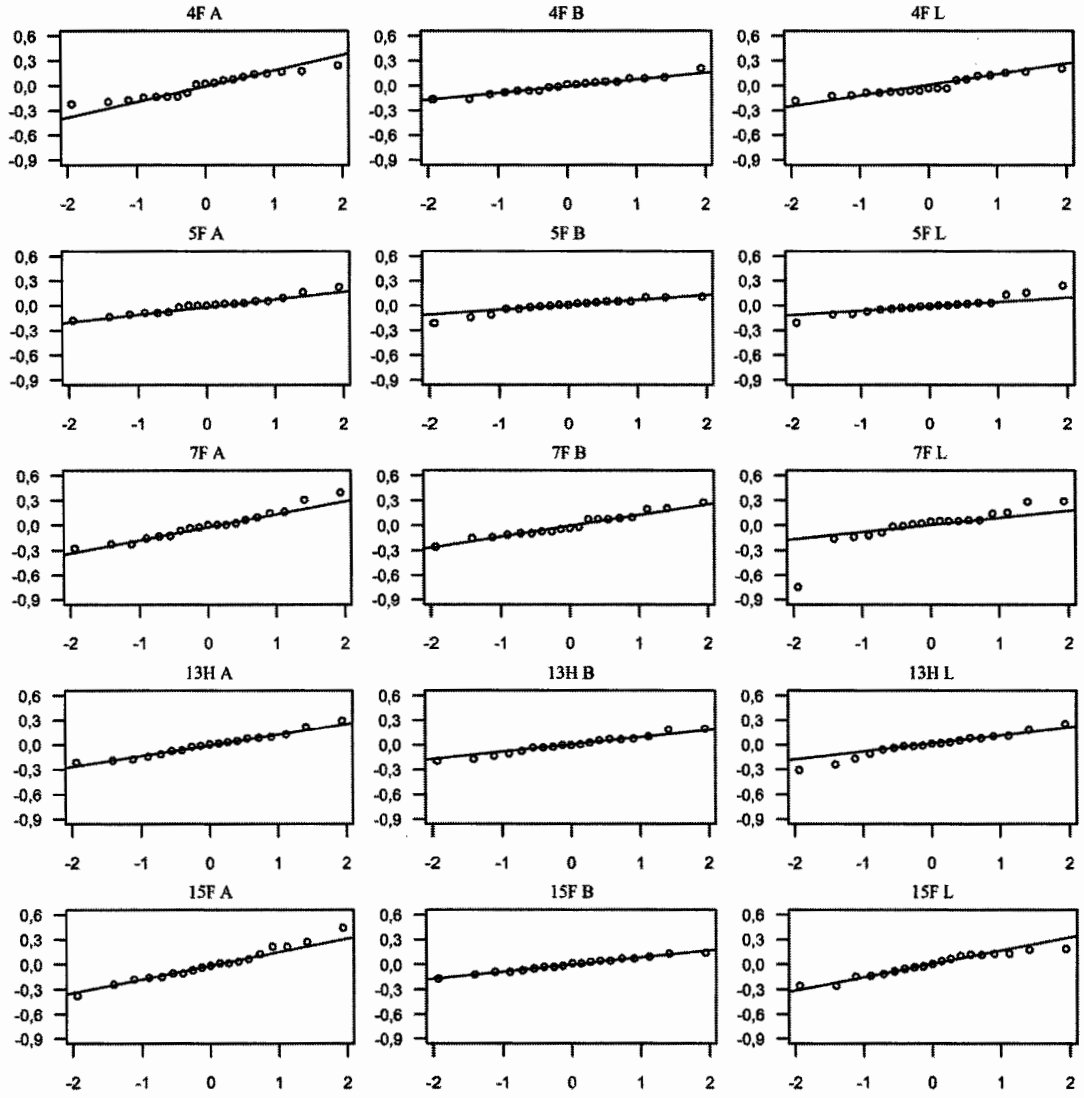


Figure 4.31 : Diagrammes quantile-quantile des séries de 4F, 5F, 7F, 13H et 15F, du 2012-12 au 2013-09.

Tableau 4.27
Statistiques descriptives des séries chronologiques de 4F, 5F, 7F, 13H et 15F, du 2012-12 au 2013-09

Série	*m*	*s*	min	max	étendue	asymétrie	aplatissement	% ±2 *s*
4F A	-1,88	0,15	-2,10	-1,63	0,47	-0,01	-1,52	100
5F A	-1,50	0,10	-1,68	-1,28	0,40	0,30	-0,30	95
7F A	-1,54	0,18	-1,81	-1,14	0,67	0,50	-0,38	95
13H A	-1,41	0,14	-1,62	-1,11	0,51	0,28	-0,65	95
15F A	-1,92	0,20	-2,29	-1,47	0,82	0,36	-0,30	95
4F B	-1,08	0,09	-1,24	-0,87	0,37	0,13	-0,47	95
5F B	-0,90	0,08	-1,11	-0,79	0,32	-0,92	0,25	95
7F B	-0,90	0,14	-1,15	-0,62	0,53	0,28	-0,87	95
13H B	-0,89	0,11	-1,08	-0,69	0,39	-0,02	-0,75	100
15F B	-1,04	0,08	-1,21	-0,90	0,31	-0,12	-0,93	100
4F L	-2,10	0,11	-2,28	-1,89	0,39	0,40	-1,26	100
5F L	-1,95	0,10	-2,16	-1,71	0,45	0,47	0,52	89
7F L	-1,85	0,22	-2,60	-1,56	1,04	-1,87	4,64	95
13H L	-2,10	0,14	-2,41	-1,85	0,56	-0,45	-0,18	95
15F L	-1,96	0,14	-2,21	-1,77	0,44	-0,38	-1,11	100

% ±2 *s* : pourcentage des valeurs à plus ou moins 2 écarts types de la moyenne

4.6.2 Modélisation AMMI

Des 15 séries de cette période, 9 ne sont que du bruit blanc. Les modèles décrivant le mieux les 6 autres séries sont présentés dans le tableau 4.28. Toutes les séries respectent les conditions d'utilisation de la modélisation AMMI, comme le montrent les résultats des tests diagnostiques en annexe A.

Tableau 4.28
Modèles AMMI, des séries chronologiques de 4F et 5F, du 2012-12 au 2013-09

Série	AMMI	Coefficients	Erreurs types
5F L	0,0,1	0,85	0,22
7F A	0,1,1	-0,55	0,21
7F L	1,0,0	0,62	0,28
13H A	0,0,1	-0,91	0,18
13H B	1,0,0	-0,38	0,21
15F A	1,1,0	-0,70	0,16

L'ordre maximal des modèles est 1 et aucun modèle ne combine de coefficient autorégressif et de moyenne mobile. Sans surprise, les 2 séries devant être intégrées (7F A et 15F A) sont les deux séries ayant la pente la plus forte, ce qui confirme que la moyenne de ces séries n'est pas temporellement stable. Les coefficients obtenus ici sont d'une ampleur assez élevée, allant en valeur absolue de 0,38 à 0,91, ce en quoi ils ressemblent beaucoup aux coefficients obtenus à la section 4.2.2, où les données sont obtenues en fonction du nombre de candidats évalués. Les coefficients des deux sections précédentes (4.4.2 et 4.5.2), eux, sont d'une ampleur plus faible, tournant autour de ±0,40. Les 2 examinateurs 7F et 13H ont chacun 2 séries ayant un modèle non nul, mais, dans les deux cas, les 2 modèles sont assez différents. Pour 7F, une série a un modèle différencié ayant une composante de moyenne mobile ayant un coefficient négatif et l'autre série a une composante autorégressive ayant un coefficient positif. Pour 13H, les deux modèles sont également distincts, ce qui est peu compatible avec la présence de mécanismes psychologiques sous-jacents communs à l'expression des niveaux de sévérité d'un même examinateur.

4.6.3 Corrélations croisées intraindividuelles

Le tableau 4.29 liste les coefficients de corrélation croisés intraindividuelles des examinateurs 4F, 5F, 7F, 13H et 15F aux délais -1, 0 et +1.

Tableau 4.29
Corrélations croisées intraindividuelles aux délais -1, 0 et +1 pour les examinateurs 4F, 5F, 7F, 13H et 15F, du 2012-12 au 2013-09

	Délai		
	-1	0	+1
4F : A et B	0,37	0,19	-0,12
4F : A et L	-0,01	0,00	0,07
4F : B et L	-0,16	0,44	0,02
5F : A et B	-0,05	0,22	-0,09
5F : A et L	-0,21	0,01	-0,09
5F : B et L	-0,33	0,29	0,11
7F : A et B	0,30	0,43	0,12
7F : A et L	-0,17	0,11	0,11
7F : B et L	-0,26	0,35	-0,10
13H : A et B	-0,40	0,02	0,06
13H : A et L	0,09	0,01	0,40
13H : B et L	-0,25	0,16	0,12
15F : A et B	-0,06	0,26	-0,12
15F : A et L	-0,25	-0,27	0,10
15F : B et L	-0,17	-0,03	-0,44

Ces résultats sont très surprenants. Non seulement les corrélations au délai 0 sont très faibles (moyenne de la valeur absolue des 15 coefficients = 0,19 et moyenne des coefficients = 0,15), mais, dans 9 cas sur 15, le coefficient au délai -1 est supérieur, en valeur absolue, au coefficient au délai 0. Si 4 des 5 coefficients entre A et B ont une valeur similaire aux valeurs observées dans les autres ensembles de données (voir figure 4.16 et tableaux 4.17, 4.21 et 4.25), le coefficient entre A et B est inférieur, en valeur absolue, à l'un ou l'autre des deux autres coefficients pour 4 des 5 examinateurs, ce qui est contraire aux attentes. Le faible nombre de temps de mesure (19) fait en sorte que ces corrélations sont calculées à partir de 19 paires de valeurs, au délai 0, ou de 18 aux délais -1 et +1. Cela peut expliquer le caractère surprenant des coefficients, mais le nombre de paires de valeurs est tout de même suffisamment élevé pour qu'il ne soit pas possible de simplement ignorer ces résultats.

4.6.4 Corrélations croisées interindividuelles

Le tableau 4.30 montre le nombre de candidats évalués conjointement par les 10 paires possibles d'examinateurs différents.

Tableau 4.30
Nombre de candidats conjointement évalués par les examinateurs 4F, 5F, 7F, 13H et 15F, du 2012-12 au 2013-09

A travaillé avec	5F	7F	13H	15F
4F	271	41	124	24
5F		93	148	100
7F			46	20
13H				36

Les 10 dyades se répartissent naturellement en deux groupes : les examinateurs ayant souvent travaillé ensemble (4F et 5F/13H ; 5F et 7F/13H/15F) et ceux ayant rarement travaillé ensemble (4F et 7F/15F ; 7F et 13H/15F ; 13H et 15F). Il serait donc plausible que les corrélations croisées des séries des examinateurs ayant souvent travaillé ensemble soient plus élevées, en valeur absolue, que les corrélations croisées des séries des examinateurs ayant rarement travaillé conjointement. Le tableau 4.31 présente les coefficients de corrélation croisée des 10 paires d'examinateurs.

Tableau 4.31
Corrélations croisées* interindividuelles aux délais -1, 0 et +1 entre les examinateurs 4F, 5F, 7F, 13H et 15F, du 2012-12 au 2013-09

	A			B			L		
	-1	0	+1	-1	0	+1	-1	0	+1
4F / 5F	*-0,16*	*-0,44*	*0,03*	*0,11*	*-0,45*	*-0,06*	*-0,02*	*-0,30*	*-0,12*
4F / 7F	0,03	-0,09	-0,04	0,12	0,64	-0,15	0,04	0,35	-0,14
4F / 13H	*0,34*	*0,01*	*-0,12*	*-0,30*	*-0,36*	*0,23*	*0,20*	*-0,62*	*-0,15*
4F / 15F	0,18	0,33	-0,29	0,06	0,09	-0,27	-0,08	0,23	0,51
5F / 7F	*-0,37*	*-0,08*	*0,41*	*-0,44*	*-0,48*	*0,02*	*-0,12*	*-0,35*	*0,26*
5F / 13H	*-0,35*	*-0,26*	*0,25*	*0,31*	*-0,20*	*0,00*	*0,17*	*0,14*	*-0,16*
5F / 15F	*0,13*	*-0,41*	*0,19*	*0,17*	*0,29*	*0,13*	*0,02*	*-0,12*	*-0,07*
7F / 13H	-0,20	0,19	0,22	-0,14	-0,52	0,12	0,07	-0,52	0,19
7F / 15F	0,07	-0,49	0,32	-0,17	-0,33	-0,09	0,35	0,00	0,01
13H / 15F	-0,50	0,22	0,27	0,08	-0,26	-0,01	-0,05	-0,38	-0,32

*Les coefficients des paires ayant souvent travaillé ensemble sont en italique

Clairement, les corrélations ne suivent pas le profil attendu. Les moyennes, en valeur absolue, des corrélations des paires ayant souvent travaillé ensemble sont presque identiques aux moyennes des paires ayant rarement travaillé ensemble (respectivement, pour A, B et L, au délai 0 : 0,24 contre 0,26 ; 0,36 contre 0,37 et 0,31 contre 0,30). De même, si l'on considère les 90 coefficients présentés dans le tableau 4.24, 26 sont égaux ou supérieurs à 0,30 en valeur absolue. De ces 26 coefficients, 13 appartiennent à des paires ayant souvent travaillé conjointement et 13 à des paires ayant rarement travaillé ensemble. Quelle que soit l'analyse utilisée, il n'y a aucune différence entre les corrélations des deux groupes de paires d'examinateurs. Lorsque l'ensemble des coefficients sont étudiés, il semble y avoir beaucoup de bruit, surtout considérant que ces corrélations sont calculées à partir de 19 paires de données (délai 0) ou 18 (délai ±1). Par exemple, pour A, la corrélation la plus forte est la celle au délai -1, entre 13H et 15F (r = -0,50). Or, ces examinateurs ont travaillé ensemble 36 fois, pour une moyenne de candidats évalués par temps de

mesure inférieure à 2. Qui plus est, toujours au délai -1, la corrélation B entre ces deux examinateurs est égale 0,08, ce qui est négligeable. La même chose se produit entre 4F et 15F, pour la note L, au délai +1 (r = 0,51). Ces deux examinateurs n'ont travaillé conjointement que 24 fois en 19 temps de mesure. Malgré cette quasi-absence de collaboration, cette corrélation est la troisième plus forte parmi les 30 corrélations L. Un dernier exemple d'incongruité : la plus forte corrélation, parmi les 90 du tableau 4.24, est la corrélation B, au délai 0, entre 4F et 7F, deux examinatrices ayant travaillé ensemble 41 fois (r = 0,64). Mais ces 41 collaborations sont toutes survenues entre le 9^{e} et le 17^{e} temps de mesure. Il est donc impossible que ces examinatrices se soient influencées ou concertées d'une manière ou d'une autre, puisque celles-ci n'ont travaillé ensemble que lors de 9 temps sur 19.

Il y a néanmoins certains patrons de corrélations vraisemblables qui correspondent aux attentes. La dyade ayant le plus souvent travaillé ensemble, 4F et 5F, ont ainsi des corrélations qui respectent la logique : pour A, B et L, les corrélations au délai 0 sont supérieures aux corrélations aux délais ±1, elles sont négatives pour les 3 notes et elles ont une étendue étroite (de -0,45 à -0,30). Les corrélations aux délais ±1 sont toutes très faibles et proches de 0. Mais c'est le seul exemple où les corrélations entre les niveaux de sévérité de deux examinateurs suivent systématiquement le même patron. Dans tous les autres cas, il n'est pas possible d'affirmer avec certitude que les corrélations observées ne sont pas le résultat du hasard.

4.6.5 Comparaisons débutants et expérimentés

Cette période comprend deux examinateurs débutants ayant commencé à travailler en décembre 2012 : 13H et 15F. L'examinateur 13H a évalué 111 candidats durant les 5 premiers temps de mesure, soit la période durant laquelle il peut être considéré débutant, suite au seuil préalablement établi de 100 candidats évalués. L'examinatrice 15F, elle, a évalué 114 candidats durant les 12 premiers temps de mesure. Nous

étudierons d'abord 13H, en ne considérant que les 5 premiers temps de mesure sur les 19 que compte la période complète. Pour les séries A, 13H a la moyenne la plus élevée, la valeur maximale la plus élevée parmi les séries des 5 examinateurs, l'étendue la plus élevée et la variance la plus élevée, ex aequo avec 2 autres examinatrices. Pour les séries B, c'est tout le contraire. 13H a la plus faible variance et son niveau de sévérité est « dans le milieu », ni plus sévère ni plus clément que celui de ses collègues. Il n'y a rien à signaler pour la série L de 13H, si ce n'est une tendance linéaire locale négative assez importante pour les 5 premiers temps de mesure.

Pour l'évolution des séries de l'examinatrice 15F, seuls les 12 premiers temps de mesure sont considérés. Pour les séries A, le niveau de sévérité de 15F se démarque clairement du niveau de ses collègues. La série 15F a la plus grande variance, la plus grande étendue, la valeur minimale parmi les 5 séries A et elle a la 2^{e} moyenne la plus basse. La différence pour la variance est particulièrement impressionnante, la série 15F A ayant une variance 2,5 fois plus grande que la série ayant la 2^{e} plus grande variance (0,05 par rapport à 0,02). Pour les séries B, la série 15F a la 2^{e} plus grande variance et étendue, mais ce sont là les seuls résultats particuliers. Quant aux séries L, la série 15F a la 2^{e} plus grande variance.

Il ressort de l'étude du niveau de sévérité de 13H et 15F que, pour les séries A, plusieurs résultats suggèrent la présence d'une période d'ajustement du jugement évaluatif pour ces examinateurs débutants, la comparaison entre l'évolution du niveau de sévérité des 2 examinateurs débutants et des 3 examinatrices expérimentées ayant révélé des différences notables. Ce n'est pas le cas pour les séries B ou L, pour lesquels les différences sont rares.

4.7 Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du temps chronologique, du 2013-11 au 2014-04

4.7.1 Représentation graphique et description

Les figures 4.32 à 4.34 montrent les graphiques chronologiques des séries chronologiques des examinateurs 12F, 13H, 17F, 18F et 20F, du 2013-11 au 2014-04. Chaque temps de mesure représente ⅓ de mois et il y a au total 18 temps. Durant cette période, 12F a évalué 95 candidats, 13H 575, 17F 217, 18F 92 et 20F en a évalué 361, pour des moyennes respectives, par temps de mesure, de 5, 32, 12, 5 et 20 candidats évalués. Les estimations de niveau de sévérité, pour chaque note, sont communes pour les 5 examinateurs et leurs niveaux de sévérité peuvent donc être directement comparés. Chaque figure montre 5 graphiques superposés, où chaque graphique est consacré à la série d'un examinateur. Toutes les ordonnées d'une même figure ont la même échelle, et les ordonnées des 3 figures ont la même étendue, soit 1,80 logit, ce qui permet la comparaison des fluctuations pour tous les graphiques de ces 3 figures.

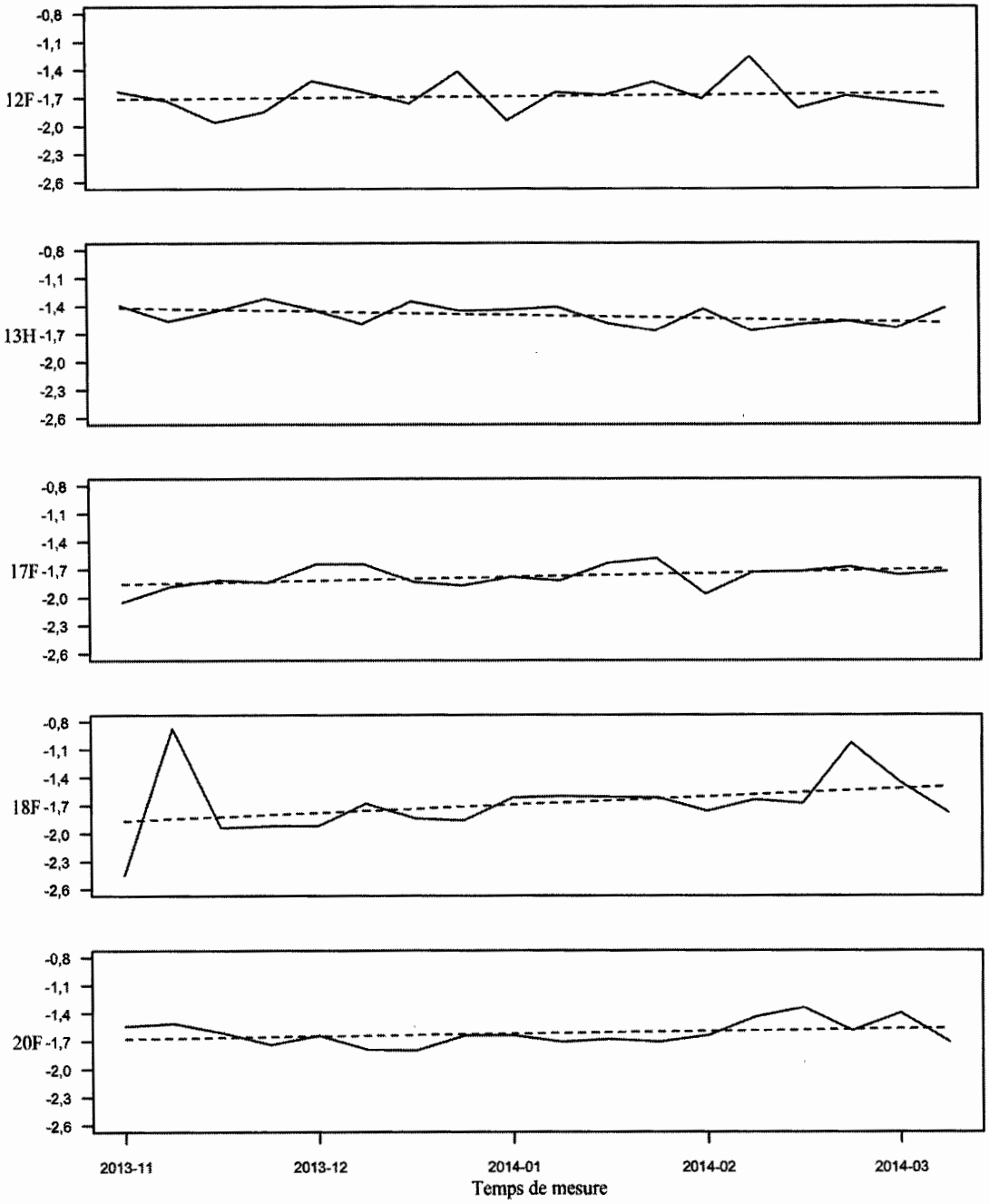


Figure 4.32 : Niveaux de sévérité A des examinateurs 12F, 13H, 17F, 18F et 20F du 2013-11 au 2014-04. L'ordonnée est en logit.

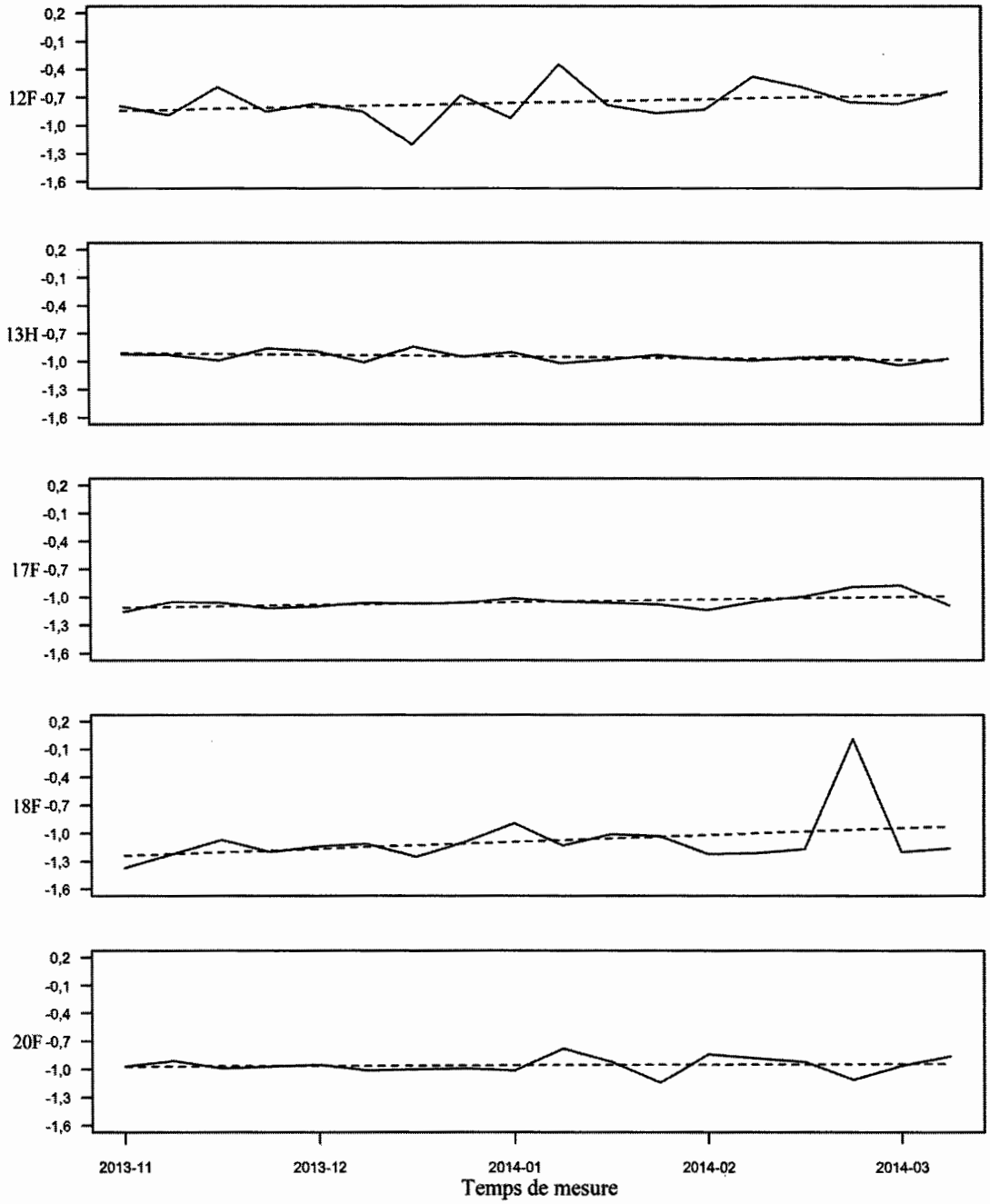


Figure 4.33 : Niveaux de sévérité B des examinateurs 12F, 13H, 17F, 18F et 20F du 2013-11 au 2014-04. L'ordonnée est en logit.

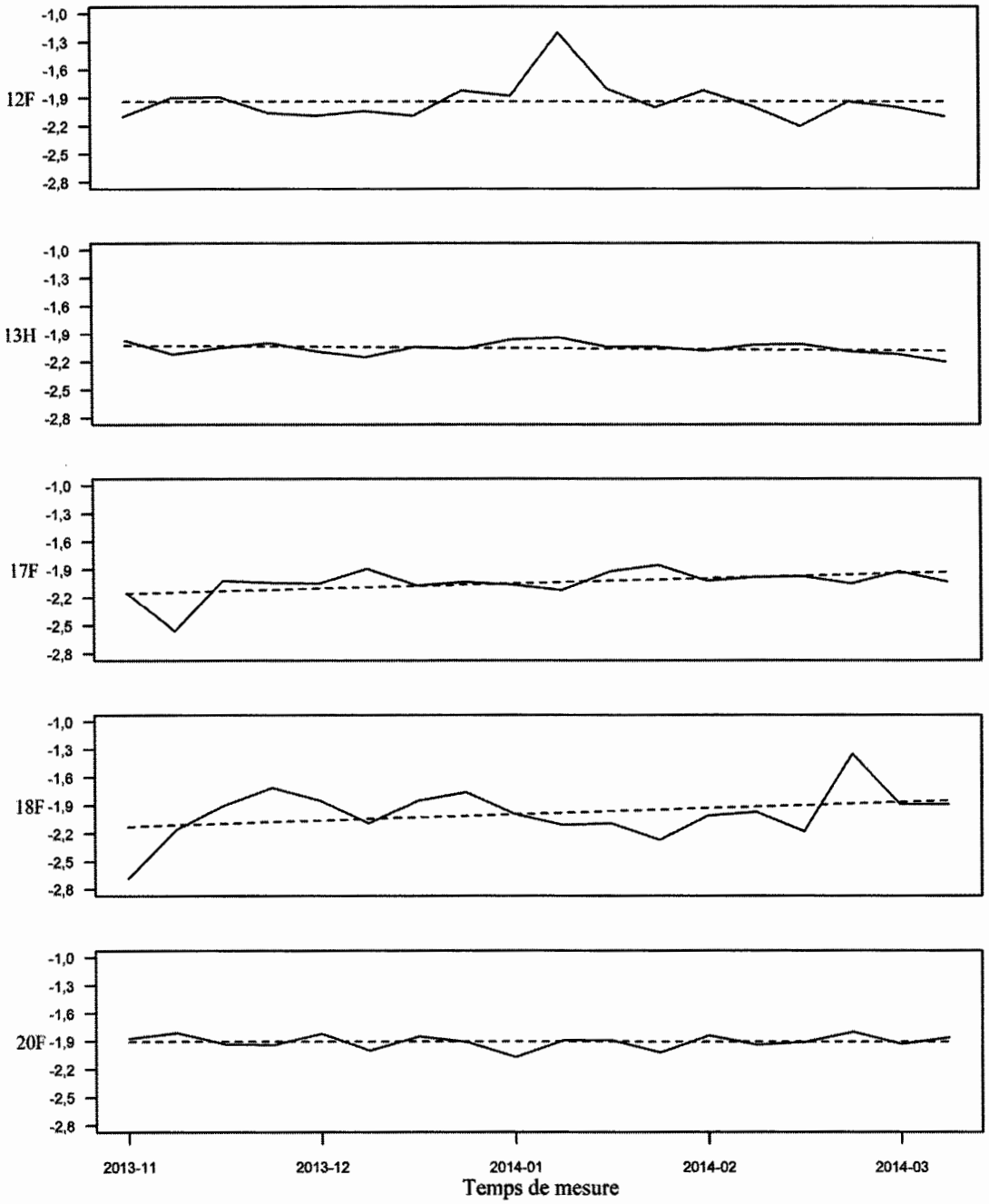


Figure 4.34 : Niveaux de sévérité L des examinateurs 12F, 13H, 17F, 18F et 20F du 2013-11 au 2014-04. L'ordonnée est en logit.

Ces 15 séries chronologiques contiennent 7 valeurs extrêmes, qui s'expliquent toutes par un faible nombre de candidats évalués lors d'un temps de mesure. Aux temps 1 des séries A et L et 16 des séries A, B et L, l'examinatrice 18F n'a évalué qu'un seul candidat, tout comme l'examinatrice 12F au temps 10 de la série L. L'examinatrice 18F a évalué 2 candidats au temps 2 de la série A, ce qui est également peu. Ces valeurs extrêmes sont également responsables de la présence des deux seules tendances linéaires locales importantes, soit la tendance descendante de janvier à février 2014 de la série 12F L et la tendance ascendante des 4 premiers temps de mesure de la série 18F L. De même, le seul plateau de ces 15 séries se trouve dans la série 18F L, du début janvier à la mi-février 2014, période durant laquelle le niveau de sévérité se maintient très légèrement sous la tendance linéaire globale. Les valeurs extrêmes susmentionnées sont vraisemblablement responsables de la présence de certaines tendances linéaires globales, 4 séries ayant une tendance linéaire globale suffisamment importante pour que la différence entre le premier et le dernier temps de mesure soit égale ou supérieure à 0,20 logit en valeur absolue : les 3 séries de 18F (différence de 0,39 ; 0,32 et 0,28 logit) et la série 17F L, dont la différence équivaut à 0,24 logit. Les autres séries sont à peu près stables.

La volatilité de l'évolution des niveaux de sévérité semble directement en lien avec le nombre de candidats évalués au cours de la période. L'examinateur ayant évalué le plus grand nombre de candidats, 13H, a la série ayant la plus faible variance, et ce pour les 3 notes. Inversement, les 2 examinatrices ayant évalué le moins de candidats, 12F et 18F, ont les séries ayant les variances les plus élevées, et ce pour les 3 notes. Bien sûr, ce résultat est également imputable, en partie, aux valeurs extrêmes identifiées ci-dessus, mais l'examen visuel des graphiques chronologiques révèle que même en l'absence de ces valeurs extrêmes, la variance des séries de ces examinatrices serait plus grande. Pour terminer, l'ordre de sévérité entre les examinateurs est assez stable. Pour chacune des 3 notes, l'examinateur le plus sévère

en moyenne est plus sévère que les 4 autres examinateurs 82 % du temps. Le deuxième examinateur le plus sévère l'est 65 % du temps, le troisième 68 % et le quatrième est plus sévère que son collègue 67 % du temps. Ces pourcentages représentent la moyenne des pourcentages propres à chacun des 3 ensembles de données. Finalement, lorsque les données de chaque série sont considérées sans égard à l'ordre temporel, un peu plus de la moitié sont normalement distribuées. Les séries 12F L, 17F B et L et les 3 séries de 18F sont plus ou moins anormalement distribuées, comme en font foi les statistiques descriptives et les diagrammes quantile-quantile de la figure 4.35 et du tableau 4.32. Il faut toutefois être prudent avec ces diagrammes, car les séries de cette période n'ont que 18 temps de mesure, ce qui est peu pour juger de l'adéquation à une distribution théorique.

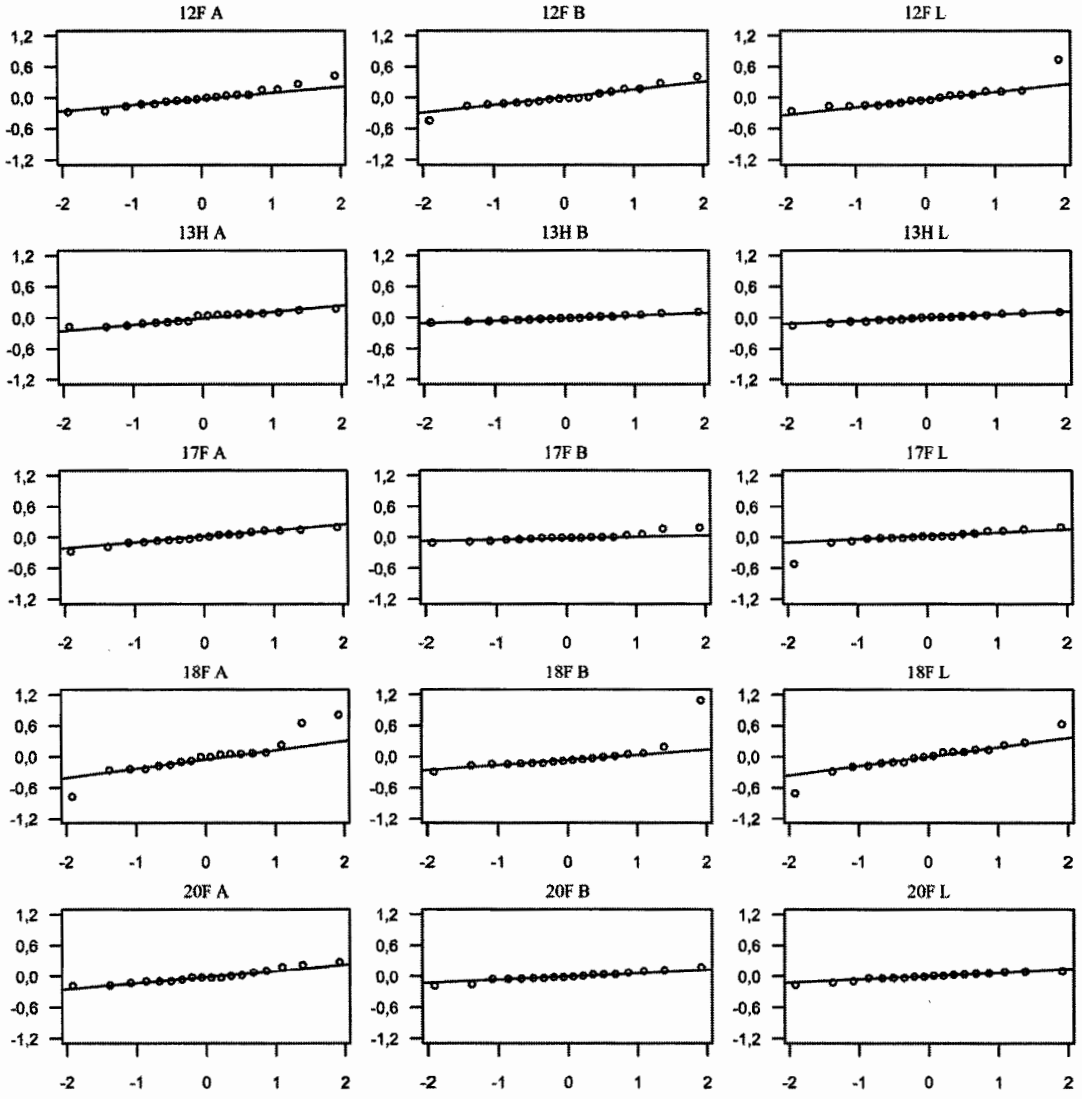


Figure 4.35 : Diagrammes quantile-quantile des séries de 12F, 13H, 17F, 18F et 20F du 2013-11 au 2014-04

Tableau 4.32
Statistiques descriptives des séries chronologiques de 12F, 13H, 17F, 18F et 20F du 2013-11 au 2014-04

Série	*m*	*s*	min	max	étendue	asymétrie	aplatissement	% ±2 *s*
12F A	-1,68	0,17	-1,96	-1,26	0,70	0,56	-0,05	94
13H A	-1,50	0,11	-1,67	-1,32	0,35	-0,14	-1,44	100
17F A	-1,77	0,12	-2,05	-1,58	0,47	-0,41	-0,56	94
18F A	-1,68	0,34	-2,45	-0,87	1,58	0,44	0,92	89
20F A	-1,62	0,13	-1,80	-1,34	0,46	0,62	-0,64	94
12F B	-0,76	0,19	-1,20	-0,35	0,85	0,00	0,39	89
13H B	-0,95	0,05	-1,04	-0,84	0,20	0,35	-0,77	94
17F B	-1,05	0,07	-1,16	-0,87	0,29	1,04	0,60	89
18F B	-1,08	0,29	-1,37	0,01	1,38	2,82	7,89	94
20F B	-0,96	0,09	-1,14	-0,78	0,36	-0,15	-0,25	94
12F L	-1,94	0,22	-2,20	-1,20	1,00	2,04	4,69	94
13H L	-2,05	0,07	-2,20	-1,94	0,26	-0,25	-0,66	94
17F L	-2,04	0,15	-2,56	-1,85	0,71	-2,06	5,06	94
18F L	-1,99	0,27	-2,69	-1,35	1,34	-0,24	1,21	89
20F L	-1,90	0,07	-2,07	-1,80	0,27	-0,55	-0,51	94

% ±2 *s* : pourcentage des valeurs à plus ou moins 2 écarts types de la moyenne

Les statistiques descriptives de ce tableau confirment les résultats de l'examen visuel des graphiques chronologiques. Les 2 examinatrices ayant évalué le moins de candidats ont les séries ayant la plus grande variance, la plus grande étendue et deux des trois valeurs les plus élevées d'asymétrie et d'aplatissement. Par ailleurs, les examinateurs ont une volatilité constante, comme le montre la variance. L'examinateur 13H a la variance la plus basse pour les 3 notes et l'examinatrice 20F, qui a évalué le deuxième plus grand nombre de candidats au cours de cette période, a une variance aussi très basse. Précisons, pour terminer, que les valeurs de chaque série sont regroupées autour de la moyenne, 11 séries sur 15 ayant 94 % ou 100 % des valeurs à ±2 écarts types de la moyenne.

4.7.2 Modélisation AMMI

Le modèle nul du bruit blanc (0,0,0) est le meilleur modèle pour 9 des 15 séries chronologiques de cette période. Les modèles les plus appropriés pour les 6 autres séries se trouvent dans le tableau 4.33. Toutes les séries respectent les conditions d'utilisation de la modélisation AMMI, comme en font foi les résultats des tests diagnostiques présentés en annexe A, à l'exception de la série 17F L, probablement hétéroscédastique. La modélisation AMMI a néanmoins été utilisée pour cette série, et ce pour les raisons énumérées à la section 4.2.2.

Tableau 4.33
Modèles AMMI, des séries chronologiques de 12F, 13H, 17F, 18F et 20F du 2013-11 au 2014-04

Série	AMMI	Coefficients*	Erreurs types
13H A	0,1,1	-0,77	0,16
17F B	0,1,0	-	-
17F L	0,1,1	-0,71	0,20
20F A	1,0,0	0,37	0,22
20F B	2,0,0	-0,26 ; -0,55	0,20 ; 0,18
20F L	3,0,0	-0,25 ; 0,08 ; 0,63	0,17 ; 0,19 ; 0,18

*Les coefficients suivent l'ordre $A_1...Ap$, $MM_1...MMq$ et sont séparés par un « ; »

Trois résultats retiennent l'attention. La présence d'un modèle d'ordre 3 (20F L), le seul de tous les modèles pour les niveaux de sévérité obtenus en fonction du temps chronologique. Le deuxième résultat est que les 3 séries de l'examinatrice ont un modèle avec une composante autorégressive, ce qui représente le seul cas où, pour les niveaux de sévérité obtenus selon le temps chronologique, les 3 séries d'un même examinateur ont un modèle non nul. Le troisième résultat est négatif, en ce sens où les séries 18F A et L, qui ont toutes deux des valeurs extrêmes au temps 1, n'ont pas de composante de moyenne mobile. En d'autres termes, les valeurs extrêmes de ces séries au temps 1 n'influencent pas, statistiquement, le niveau de sévérité aux temps

subséquents, ce qui est conceptuellement plausible, puisque ces valeurs extrêmes sont dues, rappelons-nous, au fait que 18F n'a évalué qu'un seul candidat. Sinon, les coefficients obtenus sont similaires par l'étendue et l'ampleur aux coefficients obtenus avec les autres ensembles de données et les modèles ont une composante autorégressive ou de moyenne mobile, mais pas les deux en même temps. Finalement, puisque les temps de mesure de cet ensemble de données correspondent à ⅓ de mois, notons que la dépendance temporelle du niveau de sévérité va de ~10 à ~30 jours (ordre 1 à 3), ce qui correspond bien à l'échelle de temps des dépendances observées pour les 3 ensembles de données précédents, soit de ~15 à ~30 jours (ordre 1 ou 2), puisque ces ensembles de données ont des temps de mesure correspondant à ½ mois. Il y a donc une constance à ce sujet à travers ces 4 ensembles de données.

4.7.3 Corrélations croisées intraindividuelles

Le tableau 4.34 énumère les coefficients de corrélation croisée intraindividuelles des examinateurs 12F, 13H, 17F, 18F et 20F aux délais -1, 0 et +1.

Tableau 4.34
Corrélations croisées intraindividuelles aux délais -1, 0 et +1 pour les examinateurs 12F, 13H, 17F, 18F et 20F du 2013-11 au 2014-04

	Délai		
	-1	0	+1
12F : A et B	-0,25	0,27	-0,36
12F : A et L	-0,35	0,08	-0,01
12F : B et L	-0,28	0,46	0,09
13H : A et B	-0,01	0,39	-0,16
13H : A et L	0,05	0,05	-0,28
13H : B et L	0,01	0,24	-0,09
17F : A et B	-0,34	0,09	-0,04
17F : A et L	0,29	0,38	-0,01
17F : B et L	0,06	-0,18	-0,24
18F : A et B	0,03	0,52	0,04
18F : A et L	0,08	0,38	-0,37
18F : B et L	0,01	0,62	-0,11
20F : A et B	0,07	0,25	-0,01
20F : A et L	0,20	0,26	0,18
20F : B et L	0,34	0,18	-0,48

À l'instar des corrélations obtenues à partir des données du 2012-12 au 2013-09, les corrélations de cet ensemble de données sont inattendues. Le seul examinateur ayant, pour ses 3 paires de variables, la corrélation la plus élevée, en valeur absolue, entre A et B au délai 0 est 13H. Les 4 autres examinatrices ont une corrélation plus élevée pour l'une des deux autres paires de variables, A et L (17F et 20F) ou B et L (12F et 18F). L'examinatrice 18F a bien une corrélation assez élevée entre A et B, mais sa corrélation entre B et L l'est encore plus, bien que, logiquement, les niveaux de sévérité A et B devraient avoir davantage en commun que les niveaux B et L. Une fois de plus, la moyenne des 15 corrélations au délai -1 a une moyenne très proche de 0, soit -0,01, et la moyenne au délai +1 est également très faible, -0,12. Les corrélations au délai 0, elles, ont une moyenne de 0,27, ce qui est similaire aux corrélations observées au délai 0 pour les autres ensembles de données.

4.7.4 Corrélations croisées interindividuelles

Le tableau 4.35 montre le nombre de candidats évalués conjointement par les 10 paires possibles d'examinateurs différents.

Tableau 4.35
Corrélations croisées intraindividuelles aux délais -1, 0 et +1 pour les examinateurs 12F, 13H, 17F, 18F et 20F du 2013-11 au 2014-04

A travaillé avec	13H	17F	18F	20F
12F	85	0	0	3
13H		148	70	192
17F			0	55
18F				6

Les dyades se divisent en deux groupes : les dyades n'ayant jamais ou très peu travaillé ensemble (12F et 17F/18F/20F ; 17F et 18F ; 18F et 20F) et les dyades ayant parfois ou souvent travaillé de concert (12F et 13H ; 13H et 17F/18F/20F ; 17F et 20F). Le tableau 4.36 suit et contient les corrélations croisées interindividuelles entre les examinateurs 12F, 13H, 17F, 18F et 20F.

Tableau 4.36
Corrélations croisées* interindividuelles aux délais -1, 0 et +1 entre les examinateurs 12F, 13H, 17F, 18F et 20F du 2013-11 au 2014-04

	A			B			L		
	-1	0	+1	-1	0	+1	-1	0	+1
12F / 13H	*-0,29*	*-0,49*	*0,47*	*0,05*	*-0,70*	*0,20*	*0,34*	*0,37*	*0,41*
12F / 17F	0,04	0,17	-0,09	0,10	-0,07	0,09	0,32	-0,18	-0,27
12F / 18F	0,07	-0,02	-0,28	0,30	0,02	0,00	-0,28	0,01	-0,16
12F / 20F	0,35	0,27	-0,31	0,20	0,50	0,10	-0,15	-0,33	-0,38
13H / 17F	*0,09*	*-0,47*	*0,18*	*0,30*	*0,00*	*-0,20*	*-0,26*	*0,06*	*-0,03*
13H / 18H	*-0,01*	*-0,37*	*0,16*	*-0,02*	*-0,04*	*-0,40*	*-0,11*	*-0,31*	*0,00*
13H / 20F	*0,14*	*-0,42*	*0,02*	*-0,10*	*-0,30*	*-0,05*	*0,08*	*-0,44*	*-0,45*
17F / 18F	-0,21	0,34	0,07	0,10	0,30	-0,08	0,06	0,05	0,44
17F / 20F	*-0,28*	*-0,18*	*0,15*	*0,02*	*-0,20*	*0,30*	*-0,13*	*-0,38*	*0,17*
18F / 20F	0,08	0,09	0,32	0,01	-0,40	0,50	-0,43	-0,01	0,21

*Les coefficients des paires ayant souvent travaillé ensemble sont en italique

Ces résultats sont intéressants et contrastent avec les résultats équivalents des données de la période du 2012-12 au 2013-09, mais il ne faut pas oublier que ces corrélations sont obtenues à partir de peu de données, 18 paires de données pour le délai 0 et 17 pour les délais ±1. Dans l'ensemble, les corrélations au délai 0 respectent la logique attendue pour les notes A et L. La moyenne des valeurs absolues des corrélations des 5 dyades ayant travaillé ensemble étant de 0,39 pour A et 0,31 pour L, alors qu'elle n'est que de 0,18 et 0,12 pour les dyades n'ayant jamais travaillé ensemble[20]. Pour les notes B, le constat est mitigé, les corrélations au délai 0 étant, en valeur absolue, égales entre les deux groupes d'examinateurs, soit 0,25 pour les dyades ayant travaillé ensemble et 0,26 pour celles n'ayant jamais travaillé ensemble. Par exemple, les examinatrices 12F et 20F, qui n'ont évalué conjointement que 3 candidats en 6 mois, ont une corrélation de 0,50 au délai 0 et les examinateurs 13H et 17F, qui ont travaillé ensemble 148 fois, ont une corrélation de 0. Les corrélations

[20] Afin d'alléger le texte, le terme « jamais » englobe les dyades n'ayant que très peu travaillé ensemble.

aux délais ±1 semblent également respecter la logique attendue, du moins pour les corrélations A et L. Ces corrélations, pour les dyades ayant travaillé ensemble, sont inférieures aux corrélations au délai 0 ; leur moyenne, en valeur absolue, va de 0,16 à 0,23. Or, ces moyennes, pour les dyades n'ayant jamais travaillé ensemble, vont de 0,15 à 0,29 et elles sont supérieures aux moyennes au délai 0. En résumé, il semble y avoir des différences importantes entre les corrélations A, L et B. Les premières montrent que, au délai 0, les niveaux de sévérité d'examinateurs ayant travaillé ensemble sont davantage corrélés que ne le sont les niveaux de sévérité d'examinateurs n'ayant jamais travaillé ensemble, et que ces corrélations aux délais ±1 sont inférieures aux corrélations au délai 0. Ces deux constats sont faux pour les corrélations B, qui ne révèlent aucune logique sous-jacente.

En ce qui concerne les dyades particulières, certaines montrent des patrons intéressants. Trois dyades ayant travaillé ensemble, 12F/13H, 13H/20F et 17F/20F, ont des corrélations A et B, au délai 0, d'une ampleur et d'une direction similaires. De plus, ces deux dernières dyades ont aussi une corrélation L, au délai 0, négative et d'une ampleur semblable. La dyade ayant le plus souvent travaillé ensemble au cours de cette période, 13H et 20F, a un patron de corrélations régulier. Les corrélations au délai 0 sont toutes négatives et ont une étendue limitée (de -0,44 à -0,30). Les corrélations au délai ±1 sont de faible ampleur, proches de 0, à l'exception de la corrélation L au délai +1, aussi importante que la corrélation au délai 0 de la même note. En revanche, la deuxième dyade ayant travaillé le plus souvent ensemble, 13H et 17F, a un patron de corrélations confus. La corrélation A au délai 0, de -0,47, représente un va-et-vient entre les niveaux de sévérité des deux examinateurs, qui tantôt convergent, tantôt divergent, ces niveaux ayant un écart de 0,27 logit, en moyenne. Mais la corrélation B, au délai 0, est parfaitement nulle, tandis que les corrélations aux délais ±1 sont non négligeables. Mais, pour B, les deux examinateurs ont une moyenne éloignée de seulement 0,10 logit, et les variances des deux séries

sont extrêmement faibles, respectivement 0,003 et 0,005. Leur niveau de sévérité varie très peu, d'un temps à l'autre, aussi il n'est pas très étonnant que la corrélation entre leur niveau de sévérité soit si faible, du moins au délai 0. Qu'elle soit tout de même de 0,30 au délai -1 est plus difficile à expliquer. Une autre dyade (12F et 20F) a un patron de corrélations étrange, en ce que ces examinatrices n'ont travaillé ensemble que 3 fois en 6 mois, mais la corrélation entre leur niveau de sévérité, pour A, B et L, va, en valeur absolue, de 0,27 à 0,50. Pour conclure, bien que certains patrons corrélationnels soient compatibles avec l'idée selon laquelle le niveau de sévérité d'examinateurs travaillant souvent ensemble, cette relation est loin d'être systématique ; il y a plusieurs contre-exemples.

4.8 Conclusion des résultats

Cette thèse a étudié l'évolution temporelle du niveau de sévérité d'examinateurs œuvrant dans le domaine du français langue étrangère. Après avoir vérifié que les données utilisées respectaient de manière satisfaisante les conditions du modèle Rasch à multifacettes, modèle retenu pour obtenir les valeurs estimées des niveaux de sévérité des examinateurs, l'évolution temporelle de ces niveaux a été modélisée de 6 manières différentes, selon le découpage temporel choisi. Premièrement, l'évolution temporelle a été modélisée non pas en fonction du temps réel, « chronologique », mais bien en fonction du temps de travail, de l'expérience professionnelle, l'intervalle de « temps » étant un nombre fixe de candidats évalués, 10 dans ce cas-ci. Cela a permis d'étudier, pour chaque examinateur, l'évolution de son niveau de sévérité en prenant en compte tous les candidats évalués d'octobre 2010 à avril 2014 par chacun et donc d'exploiter toutes les données disponibles. En revanche, cette modélisation fait en sorte que les évolutions temporelles du niveau de sévérité de différents examinateurs ne peuvent pas être directement comparées, puisque le découpage temporel est propre à chaque examinateur. Les 5 autres modélisations permettent de pallier ce problème. Ces 5 modélisations ont été faites en fonction du temps

chronologique : 5 périodes, allant d'une durée de 6 à 30 mois consécutifs, ont été identifiées et découpées en intervalles allant de ⅓ de mois à 3 mois selon le cas. Ces périodes ont été choisies de manière à ce qu'au moins 2 examinateurs ayant évalué des candidats à chacun des intervalles de temps aient évalué conjointement plusieurs candidats, de manière à ce que les données soient liées et que des estimations communes puissent être effectuées. La nature exploratoire et descriptive de cette thèse fait en sorte qu'il n'est pas possible de simplement synthétiser des résultats qui se résumeraient, par exemple, à quelques statistiques inférentielles. Les résultats de chacune des 6 modélisations sont nombreux et la présentation graphique de certains de ces résultats est essentielle. Il y a toutefois trois aspects importants des résultats sur lesquels nous revenons ici et qui tiennent lieu de synthèse : 1) la normalité des distributions atemporelles des niveaux de sévérité des examinateurs ; 2) les résultats des modélisations AMMI et 3) les rapports entre les étendues intraindividuelles des examinateurs à travers le temps et les étendues interindividuelles des examinateurs d'une modélisation temporelle donnée.

4.8.1 La normalité des distributions de niveaux de sévérité

Parmi les 6 modélisations temporelles retenues, la modélisation de la section 4.3, qui ne contient que 10 temps de mesure, est ici exclue, car il ne saurait être question d'adéquation à une distribution lorsqu'il n'y a que 10 données observées. Il reste donc 5 modélisations comptant en tout 78 séries chronologiques (voir les détails aux sections 4.2 et 4.4 à 4.7). De ces 78 séries chronologiques, 73 % ont été jugées approximativement normalement distribuées, pourcentage se répartissant comme suit à travers les 5 modélisations temporelles : 29/36 (81 %) ; 3/6 (50 %) ; 5/6 (83 %) ; 11/15 (73 %) et 9/15 (60 %). Les proportions sont relativement constantes pour les 5 modélisations, les pourcentages individuels allant de 50 % à 83 %. Au-delà de la normalité dans son ensemble, l'asymétrie de ces distributions est plus particulièrement intéressante, car elle renseigne quant aux tendances des

examinateurs à être plus sévère ou plus clément qu'ils ne le sont habituellement. Si l'on considère la moyenne ou la médiane de la distribution des niveaux de sévérité d'un examinateur pour tous les intervalles *t* d'une période donnée, alors le fait que le coefficient d'asymétrie de cette distribution soit faible (p. ex. compris entre -0,50 et 0,50) montre que cet examinateur est à peu près aussi souvent plus sévère que plus clément qu'il ne l'est globalement, ce qui est substantiellement intéressant. À l'inverse, les quelques distributions ayant des coefficients égaux ou supérieurs à 1 en valeur absolue (1H B et L, 5F A, 7F L, 12F L, 17F B et L ainsi que 18F B, toutes provenant de modélisation en fonction du temps chronologique) montrent, pour ces distributions, la présence de quelques valeurs extrêmes de sévérité ou de clémence pour ces examinateurs. Le faible nombre de ces distributions asymétriques permet toutefois d'affirmer que les examinateurs étudiés dans cette thèse, quelle que soit la modélisation temporelle utilisée, tendent à être aussi souvent localement cléments que sévères. Le faible nombre de plateaux observés dans les différents graphiques chronologiques présentés dans cette section est en accord avec cette idée, « plateaux » désignant une série de valeurs consécutives du niveau de sévérité situées au-dessus ou au-dessous de la tendance linéaire globale de ce niveau de sévérité.

4.8.2 Résultats des modélisations AMMI

Toujours en faisant abstraction des données de la section 4.3 à cause de leur faible nombre de temps de mesure, les résultats des modélisations AMMI révèlent une régularité intéressante. Sur les 78 séries chronologiques restantes, 30 (38 %) ont un modèle AMMI non nul, où au moins l'un des paramètres *p*, *d* ou *q* est d'ordre 1 ou supérieur. Ce pourcentage est réparti comme suit parmi les 5 modélisations temporelles : 12/36 (33 %), 3/6 (50 %), 3/6 (50 %), 6/15 (40 %) et 6/15 (40 %). Ces 5 proportions sont similaires, allant de 33 % à 50 %. De ces 30 modèles non nuls, 14 sont pour des niveaux de sévérité A, 9 pour B et 7 pour L. Les paramètres de la modélisation AMMI sont à peu près également représentés parmi ces 30 modèles

comme en fait foi la répartition suivante : 14 séries ont un coefficient p non nul, 12 un coefficient d et 18 un coefficient q[21]. Il ne semble donc pas y avoir de modèle dominant qui serait significativement plus présent pour les séries chronologiques de ces examinateurs. Finalement, remarquons que certains examinateurs sont présents dans 2, 3 ou 4 modélisations temporelles. Par exemple, 4F et 5F sont présentes dans la modélisation en fonction du temps de travail (section 4.2), dans la modélisation du 2012-06 au 2013-11 (section 4.5) et dans celle du 2012-12 au 2013-09 (section 4.6) ; 5F est, en sus, présente dans la modélisation du 2011-09 au 2013-02 (section 4.4). Or, les différents modèles AMMI obtenus pour ces examinatrices ne semblent pas concorder au-delà du hasard comme le montre le tableau 4.37.

Tableau 4.37
Comparaisons entre les modèles AMMI des examinatrices 4F et 5F pour différents ensembles de données

	4F			5F		
	A	B	L	A	B	L
Par 10 candidats	1,0,0	1,1,2	0,0,0	3,0,2	0,0,0	0,0,0
2011-09 au 2013-02	-	-	-	0,0,1	1,0,0	0,0,0
2012-06 au 2013-11	1,0,0	0,0,2	0,0,0	0,0,0	0,0,0	1,0,2
2012-12 au 2013-09	0,0,0	0,0,0	0,0,0	0,0,0	0,0,0	0,0,1

- L'examinatrice 4F n'est pas présente dans cette modélisation

Si les modèles de l'examinatrice 4F ont une certaine constance, ceux de l'examinatrice 5F changent davantage d'un ensemble de données à un autre. L'examinatrice 20F présente un cas intéressant à cet égard. Cette examinatrice a exclusivement travaillé du 2013-11 au 2014-04, ce qui coïncide avec l'une des modélisations temporelles retenues (section 4.7). Cette examinatrice est par conséquent présente dans 2 modélisations qui, chacune, comprend l'ensemble des candidats évalués par celle-ci, la différence résidant dans le découpage temporel

[21] Une même série peut avoir un modèle ayant 2 ou 3 paramètres non nuls, ce qui explique la somme supérieure à 30.

effectué : soit en en fonction du temps de travail (par 10 candidats) ou en fonction du temps chronologique du 2013-11 au 2014-04. Or, les modèles AMMI obtenus pour cette examinatrice sont très différents. Dans la modélisation par 10 candidats, 20F a un seul modèle non nul, pour les données A : 0,1,1. Au contraire, pour la modélisation du 2013-11 au 2014-04, 20F a 3 modèles AMMI non nuls : 1,0,0 pour les données A, 2,0,0 pour B et 3,0,0 pour L. Les 2 modèles A/B/L de 20F sont donc différents pour ces 2 modélisations, cela parce que les intervalles de temps sont définis différemment. Difficile avec de tels résultats de faire une inférence quant à la nature de la « sévérité » d'un examinateur.

4.8.3 Rapports entre étendues intra et interindividuelles

L'une des grandes préoccupations des organismes responsables d'évaluations à forts enjeux est de s'assurer que les examinateurs évaluant les performances de candidats ont, collectivement, des niveaux de sévérité relativement similaires et qu'il n'y a pas d'examinateurs ayant un niveau de sévérité beaucoup plus élevé ou plus faible que ses collègues. Une analyse transversale des niveaux de sévérité des examinateurs devrait révéler de faibles écarts entre les niveaux de sévérité des examinateurs d'une épreuve donnée. Cette approche repose donc sur la comparaison interindividuelle des niveaux de sévérité des examinateurs. Cette idée, communément admise par les communautés professionnelles et scientifiques, repose sur un présupposé : le niveau de sévérité d'un examinateur donné est, temporellement, suffisamment stable pour que l'idée même d'une analyse transversale et d'une comparaison interindividuelle ait un sens. Car si le niveau de sévérité des examinateurs a une étendue intraindividuelle aussi, voire plus grande que l'étendue interindividuelle, il ne sert à rien de savoir qu'à un moment t l'écart interindividuel des niveaux de sévérité n'est pas trop grand, car cela pourrait changer dramatiquement à un autre temps, la faible étendue interindividuelle pourrait n'être que l'effet du hasard d'avoir choisi de mesurer les niveaux de sévérité au temps t plutôt que $t + 1$. À moins de supposer

l'idée invraisemblable que les variations longitudinales intraindividuelles sont similaires pour tous les examinateurs d'une organisation. Supposons par exemple, au temps 1, deux examinateurs E1 et E2 ayant un niveau de sévérité respectif de 0 et 0,20 logit et, au temps 2, de 3 et 3,20 logit. Ici, l'étendue intraindividuelle est considérable (3,00 logits) et l'étendue interindividuelle négligeable (0,20 logit), mais cela n'a pas d'impact important pour l'évaluation, car la variation entre les temps 1 et 2 est égale pour chaque examinateur. Cette situation semble toutefois peu plausible et il importe de vérifier l'état des choses pour les examinateurs de cette thèse.

Le rapport des étendues intra et interindividuelles des niveaux de sévérité des examinateurs des 6 modélisations temporelles retenues dans cette thèse a été étudié de la manière suivante, et ce pour chaque ensemble de données (A, B et L). D'abord, l'étendue intraindividuelle des niveaux de sévérité d'un examinateur a été calculée, soit la valeur correspondant à la différence entre son niveau de sévérité maximal et minimal pour l'ensemble des temps de mesure de cette modélisation. L'étendue interindividuelle, elle, est égale à la valeur suivante. Pour chacun des temps de mesure d'une modélisation, la différence entre le niveau de sévérité le plus élevé et le plus bas est calculée. La plus grande de ces différences, pour l'ensemble des temps de mesure, correspond à l'écart interindividuel (qui est, en fait, l'écart interindividuel maximal, puisque c'est la plus grande étendue pour l'ensemble des temps de mesure de cette période). Prenons, pour illustrer, les données A de la modélisation du 2011-09 au 2013-02, qui a 34 temps de mesure et 2 examinateurs, 1H et 5F. L'examinateur 1H a eu, sur les 34 temps e mesure, un niveau de sévérité minimal de -1,72 logit et un niveau maximal de -0,91 logit. Son étendue intraindividuelle est donc égale à 0,81 logit. L'étendue interindividuelle, elle, est calculée ainsi : pour chacun des 34 temps de mesure, la différence entre les niveaux de sévérité de 1H et 5F est calculée, ce qui donne 34 valeurs. L'étendue interindividuelle correspond à la plus grande de ces 34

valeurs, soit l'écart maximal observé au cours de cette période entre les niveaux de sévérité de 1H et 5F[22].

Le rapport entre les étendues intra- et interindividuelles est ensuite exprimé sous forme de fraction. Une valeur égale à 1 signifie une égalité entre ces deux étendues, une valeur inférieure à 1 signifie que l'étendue intraindividuelle est inférieure à l'étendue interindividuelle et une valeur supérieure à 1 indique l'étendue intraindividuelle est plus grande que l'étendue interindividuelle. Ces rapports ont été calculés pour chaque examinateur des 6 modélisations temporelles, ce qui donne un total de 28 rapports, et ce pour chaque ensemble de données (A, B et L). Le tableau 4.38 contient les statistiques descriptives des étendues intraindividuelles, le tableau 4.39 les statistiques descriptives des étendues interindividuelles maximales et le tableau 4.40 montre les statistiques descriptives de la distribution des 28 valeurs de rapports.

Tableau 4.38
Statistiques descriptives des 28 étendues intraindividuelles pour les 6 modélisations temporelles retenues pour cette thèse

	min	25 %*	md	75 %*	max	*m*	s
A	0,28	0,47	0,69	0,84	1,58	0,28	0,47
B	0,20	0,37	0,51	0,61	1,38	0,20	0,37
L	0,17	0,43	0,69	1,11	1,90	0,17	0,43

* Respectivement 1er et 3^{e} quartile

[22] Nous avons fait le choix de retenir l'étendue interindividuelle maximale. Nous aurions pu, plutôt, calculer la moyenne des écarts interindividuels de chacun des temps de mesure, mais cela aurait nécessairement donné une valeur plus faible pour l'étendue interindividuelle, ce que nous voulions éviter.

Tableau 4.39
Statistiques descriptives des 6 étendues interindividuelles maximales pour les 6 modélisations temporelles retenues pour cette thèse

	min	*m*	max
A	0,77	0,95	1,06
B	0,45	0,67	1,12
L	0,31	0,68	0,92

Tableau 4.40
Statistiques descriptives des 28 rapports entre les étendues intraindividuelles et interindividuelles pour les 6 modélisations temporelles retenues pour cette thèse

	min	25 %*	md	75 %*	max	*m*	s
A	0,32	0,49	0,66	0,89	1,49	0,71	0,28
B	0,18	0,66	0,91	1,24	1,96	0,94	0,42
L	0,28	0,60	0,86	1,43	2,44	1,02	0,53

* Respectivement 1er et 3^{e} quartile

Rappelons tout d'abord que les 28 rapports proviennent du total des combinaisons « examinateur × modélisation temporelle » et qu'il est possible qu'un même examinateur ait, dans une modélisation temporelle, un rapport inférieur à 1 et un rapport supérieur dans une autre modélisation. C'est entre autres le cas de l'examinatrice 4F, qui a, pour les données A, un rapport de 1,08 selon la modélisation temporelle du temps de travail (10 candidats), mais un rapport de 0,53 selon la modélisation temporelle du 2010-10 au 2013-03. Cela dit, observons que les distributions des valeurs des rapports diffèrent selon l'ensemble de données, les rapports A ayant tendance à être plus faibles que les rapports B et L (surtout pour la médiane, le 3^{e} quartile et le maximum). Les étendues intraindividuelles sont donc plus petites, par rapport aux étendues interindividuelles, pour les données A. En revanche, pour les données B et L, selon que l'on considère la moyenne ou la médiane, près de la moitié des examinateurs ont une étendue intraindividuelle du niveau de sévérité au moins aussi élevée que l'étendue interindividuelle des niveaux de sévérité (médianes B = 90,1 et L = 0,86) et environ 25 % des examinateurs ont une

étendue intraindividuelle franchement supérieure à l'étendue interindividuelle (selon les valeurs du 3^{e} quartile, B = 1,24 et L = 1,43). Les rapports entre les étendues intra et interindividuelles sont particulièrement élevés pour les données L, comme le montrent les valeurs du 3^{e} quartile et maximales. Cela signifie qu'il y a environ 7 examinateurs (le quart de 28) ayant une étendue intraindividuelle de leur niveau de sévérité beaucoup plus élevée que l'étendue interindividuelle maximale. Cela est très intéressant, car cela va à l'encontre de l'idée communément admise selon laquelle les analyses transversales fondées sur des comparaisons interindividuelles des niveaux de sévérité mènent à des résultats permettant de juger de la validité du travail des examinateurs. Pour conclure, les résultats du tableau 4.38 montrent que, pour les données de cette thèse, toute analyse transversale est sujette à caution et peut mener, selon le hasard du moment, à des conclusions très différentes quant à l'étendue interindividuelle des niveaux de sévérité des examinateurs à un moment *t*. D'autre part, ces rapports entre les étendues intra et interindividuelles appuient la pertinence d'étudier l'évolution longitudinale des niveaux de sévérité des examinateurs, puisque, du moins pour les examinateurs de cette thèse, les variations intraindividuelles sont souvent aussi importantes que les variations interindividuelles.

CHAPITRE V

DISCUSSION

Ce chapitre est divisé en 5 sections, réparties en ordre croissant de considérations générales conceptuelles. La première section est consacrée à une comparaison des résultats de cette thèse aux résultats des 8 études recensées au chapitre 2, section 2.4, quant au phénomène de dérive temporelle du niveau de sévérité des examinateurs. La deuxième section revient sur les statistiques descriptives des séries chronologiques du niveau de sévérité des examinateurs, sur les différences entre les séries chronologiques des données A, B et L ainsi que sur les résultats de la modélisation AMMI. La troisième section examine les résultats concernant les comparaisons entre examinateurs débutants et expérimentés et les confronte aux résultats recensés à la section 2.3.4. La quatrième section traite d'une question méthodologique incontournable, soit le problème du choix du découpage temporel dans une étude longitudinale de l'évolution temporelle du niveau de sévérité. Finalement, la cinquième section concerne la question conceptuelle fondamentale : qu'est-ce que « la » sévérité d'un examinateur?

5.1 La dérive temporelle du niveau de sévérité

Notre conclusion, suite à la recension des 8 études présentées à la section 2.4, était qu'environ 33 % des examinateurs de ces études avait un niveau de sévérité faisant montre de dérive temporelle, cette dernière représentant une étendue intraindividuelle d'au moins 1 logit. Si nous retenons le même critère, alors 36 % des examinateurs de cette thèse fait preuve de dérive temporelle (voir les tableaux 4.13, 4.19, 4.23, 4.27 et

4.32). La concordance est presque parfaite. Une telle comparaison directe est toutefois trompeuse, en ce que ces résultats proviennent de situations d'évaluation utilisant des échelles d'appréciation différentes quant aux nombres de critères et d'échelons, ce qui a un effet sur l'étendue en logit des résultats obtenus suite aux analyses de Rasch à multifacettes. Il n'est pas possible de comparer directement des mesures en logit provenant, par exemple, d'échelles d'appréciation ayant un nombre différent d'échelons : *et ceteris paribus*, moins il y a d'échelons utilisés, plus grande sera l'étendue en logit[23] (Linacre et Wright, 1989). Une valeur de 1 logit, bien que mathématiquement identique d'une étude à une autre, n'a pas la même valeur conceptuelle et pragmatique.

Un autre élément ayant un impact important sur l'étendue des résultats, en logits, est l'étendue des niveaux d'habileté des candidats évalués. Cette thèse analyse les données provenant de candidats couvrant l'ensemble du continuum théoriquement possible de niveaux d'habileté en français, de candidats ne connaissant littéralement que le mot « Bonjour » jusqu'à des candidats francophones, ayant un diplôme universitaire de 2^{e} ou 3^{e} cycle d'une université francophone et travaillant en français dans le milieu des communications. À contrario, l'étude de Wolfe *et al.* (2007), par exemple, analyse les performances d'étudiants à un « *AP English Literature and Composition Examination* », ce qui suppose une population d'étudiants relativement homogène, tous à la fin du *high school* et inscrits à un cours d'anglais langue d'enseignement avancé. Cette comparaison directe, reposant sur une valeur absolue en logit comme seuil dichotomique pour affirmer la présence de dérive temporelle est ainsi inappropriée, car les diverses recherches recensées, de même que cette thèse, font appel à des situations d'évaluation aux éléments différents (nombre d'échelons de l'échelle d'appréciation, continuum des niveaux d'habileté des candidats, grille

[23] Par exemple, pour les données A de cette thèse, si l'on transforme les données brutes allant de 0 à 20 en données correspondant aux niveaux de compétence du CECRL, allant de 0 à 6, l'étendue des niveaux d'habileté des candidats passe de 19,38 à 30,51 logits et celle des examinateurs passe de 0,98 à 2,51 logits.

d'évaluation analytique ou holistique, modèle d'analyse *rating scale* ou à crédit partiel...).

Une comparaison plus rigoureuse et instructive consiste à utiliser plutôt le rapport entre les étendues intra et interindividuelles du niveau de sévérité des examinateurs afin de voir jusqu'à quel point le niveau de sévérité d'un examinateur donné fluctue temporellement par rapport aux fluctuations interindividuelles observées durant la période étudiée. La comparaison des résultats de cette thèse, présentés à la section 4.8.3 (voir le tableau 4.40), aux résultats des 5 études pour lesquelles ces informations sont disponibles (Congdon et McQueen, 2000 ; Davis, 2016 ; H. J. Kim, 2011 ; Lim, 2009 ; Lumley et McNamara, 1995) est très instructive. Ces 5 études ont un total de 65 examinateurs : 65 rapports entre les étendues intraindividuelles et interindividuelles des niveaux de sévérité ont donc été calculés de la même manière que les 28 rapports de cette thèse (voir sections 3.4.2.5 et 4.8.3). Le tableau 5.1 présente les statistiques descriptives de ces 65 rapports.

Tableau 5.1
Statistiques descriptives des 65 rapports entre les étendues intraindividuelles et interindividuelles pour les 5 études susmentionnées

n	min	25 %*	md	75 %*	max	*m*	s
65	0,05	0,18	0,26	0,39	0,75	0,29	0,17

* Respectivement 1er et 3^{e} quartile

La comparaison est éloquente : même si nous ne considérons que les résultats de cette thèse pour les ensembles de données A, pour lesquelles les rapports sont les plus faibles, les 28 rapports de cette thèse sont beaucoup plus élevés que les 65 rapports des 5 études recensées. Pour nos données A, le 1er quartile est de 0,49, ce qui est supérieur au 3^{e} quartile des 65 rapports des 5 études et la médiane de nos données A, 0,66, est tout juste inférieure à la valeur maximale des rapports observés dans la

littérature (0,75). Le constat est sans appel : comparées aux étendues interindividuelles du niveau de sévérité, les étendues intraindividuelles des examinateurs de cette thèse sont beaucoup plus élevées que celles des examinateurs des 5 études susmentionnées. Une explication de cette différence est possible : les 28 rapports de cette thèse proviennent de 6 modélisations temporelles ayant comme caractéristiques un nombre relativement restreint d'examinateurs (de 2 à 12) et un très grand nombre de temps de mesure (de 10 à 120). En contrepartie, les 65 rapports des 5 études mettent en jeu un nombre relativement élevé d'examinateurs (de 9 à 24), mais un nombre très faible de temps de mesure (de 3 à 8). Or, plus il y a d'examinateurs, plus grande est la probabilité qu'il y ait au moins un examinateur ayant un niveau de sévérité éloigné du niveau de ses collègues, ce qui fait qu'il y a une corrélation positive entre l'étendue interindividuelle et le nombre d'examinateurs. Inversement, plus il y a de temps de mesure, plus la probabilité augmente qu'un examinateur, à un temps *t*, ait un niveau de sévérité « extrême » et une étendue intraindividuelle plus élevée, ce qui fait qu'il y a une corrélation positive entre l'étendue intraindividuelle et le nombre de temps de mesure[24]. L'étendue intraindividuelle du niveau de sévérité de nos examinateurs est-elle relativement plus grande que celle des examinateurs des études antérieures parce que nos examinateurs diffèrent en leurs pratiques professionnelles, leur jugement, leur formation, etc.? Ou est-ce parce que, simplement, les caractéristiques structurelles des données collectées (temps de mesure et nombre d'examinateurs) sont positivement corrélées à la valeur du rapport des étendues intra et interindividuelles des niveaux de sévérité? Malheureusement, nos résultats ne permettent pas de répondre à ces questions.

Au-delà de ces considérations, quels sont les impacts pratiques de telles dérives temporelles du niveau de sévérité de ces examinateurs pour la situation d'évaluation

[24] Pour cette thèse, pour les données A, B et L, les coefficients de corrélation entre le nombre de temps de mesure et l'étendue intraindividuelle sont respectivement de 0,44 ; 0,28 et 0,60.

étudiée dans cette thèse? Répondre à cette question est complexe, car un écart en logit ne peut être converti directement en un écart sur l'échelle d'appréciation originale, allant de 0 à 20, puisque la relation entre ces deux valeurs est logistique et non linéaire. La relation est toutefois quasi linéaire entre les notes de 5 et 17 (voir annexe B), ce qui permet une approximation utile. Comme le montre le tableau 4.38, pour les notes A, les 28 étendues intraindividuelles vont de 0,28 à 1,58 logit, avec une médiane de 0,69. Cette étendue médiane, 0,69 logit, correspond à peu près à une différence de 1,5 sur l'échelle d'appréciation originale – en autant que ces notes se situent entre 5 et 17. Pour les notes B, les 28 étendues intraindividuelles vont de 0,20 à 1,38 logit, avec une médiane de 0,51. Cette dernière valeur correspond à peu près à une différence de 1,8 sur l'échelle d'appréciation originale. Finalement, pour les notes L, les 28 étendues vont de 0,17 à 1,90 logit, avec une médiane valant aussi 0,69, ce qui correspond approximativement à une différence de 1,2 sur l'échelle d'appréciation originale (voir l'annexe B pour les valeurs sur lesquelles sont fondées ces estimations).

Ainsi, selon le moment auquel une évaluation a lieu, la dérive temporelle du niveau de sévérité d'un examinateur *X* peut faire en sorte que les notes brutes accordées par cet examinateur peuvent différer, en moyenne, d'environ 1 à 2 points par rapport aux notes qu'il aurait pu accorder à un autre moment. Ces différences sont assez importantes pour avoir un effet indésirable quant à la note obtenue par un candidat et, partant, quant au niveau de compétence dans lequel se verra classé ce candidat, car chacun de ces niveaux a une étendue de 2 à 4 points sur l'échelle brute de 0 à 20[25]. Il est donc réaliste de supposer qu'un candidat qui aurait été classé au niveau 3 au temps *t* pourrait être classé au niveau 2 à un autre temps. Ces effets négatifs de la dérive temporelle du niveau de sévérité sont, en moyenne, légèrement moins importants que les effets négatifs des écarts interindividuels des niveaux de sévérité, comme le

[25] Rappelons que l'échelle d'appréciation de 0 à 20 est arrimée aux 6 niveaux de compétence du CECRL, et que ce sont ces niveaux qui intéressent les utilisateurs de ces tests.

montrent les résultats des tableaux 4.38 et 4.39, puisque les étendues intraindividuelles tendent à être inférieures à l'étendue interindividuelle maximale pour une période d'évaluation donnée. En revanche, la dérive temporelle du niveau de sévérité peut avoir des effets négatifs plus importants que les écarts interindividuels des niveaux de sévérité, puisque les valeurs maximales des étendues intraindividuelles sont supérieures aux valeurs maximales des étendues interindividuelles.

Les résultats de cette thèse montrent donc que la dérive temporelle du niveau de sévérité est un problème potentiellement aussi important que les écarts interindividuels des niveaux de sévérité, du moins pour la situation d'évaluation étudiée ici. Il est bien sûr impossible de généraliser les résultats obtenus dans cette thèse, les caractéristiques de la situation d'évaluation d'où viennent les données étant assez différentes des caractéristiques des autres situations d'évaluation : le nombre de critères d'évaluation, le nombre d'échelons de l'échelle d'appréciation, le fait que toutes les évaluations soient faites par 2 examinateurs procédant ensuite à un arbitrage représentent des particularités que l'on trouve rarement dans les autres situations d'évaluation à forts enjeux. Les résultats publiés dans les 5 études recensées susmentionnées (voir tableau 5.1) montrent que les problèmes potentiels causés par la dérive temporelle du niveau de sévérité sont beaucoup moins importants pour ces situations d'évaluation, mais nous revenons au constat selon lequel les caractéristiques des devis de collecte de données (temps de mesure et nombre d'examinateurs) peuvent très bien expliquer ces différences. Rien ne nous permet d'affirmer que ces problèmes n'apparaîtraient pas, si ces 5 études avaient aussi un nombre de temps de mesure allant de 10 à 120, comme dans cette thèse.

Une autre manière de conceptualiser la question de la dérive temporelle du niveau de sévérité est de s'intéresser aux plateaux présents dans les graphiques chronologiques

de l'évolution temporelle du niveau de sévérité, soit les périodes au cours de laquelle le niveau de sévérité d'un examinateur se maintient au-dessus ou au-dessous de la tendance linéaire globale de son niveau de sévérité. Ces plateaux représentent donc une forme de dérive temporelle du niveau de sévérité. Or, parmi les 84 séries chronologiques représentées graphiquement dans les résultats, seuls 10 plateaux ont été identifiés, ce qui est très peu. Surtout que la plupart de ces plateaux montrent un écart minime à la tendance linéaire globale, de l'ordre de 0,05 à 0,30 logit (voir les détails aux sections pertinentes des résultats). Cela montre que ces problèmes de dérive temporelle tendent à être circonscrits dans le temps ; il y a des écarts intraindividuels importants pour certains examinateurs, mais ces écarts tendent à se résorber assez rapidement.

Comme nous l'avons vu aux sections 2.3.3 et 2.3.4, l'une des questions les plus étudiées quant au niveau de sévérité des examinateurs concerne l'impact de la formation et de l'expérience sur le niveau de sévérité. Les résultats des études recensées dans ces 2 sections ne permettaient pas de statuer sur d'éventuels effets de la formation et de l'expérience sur le niveau de sévérité. Nos résultats permettent-ils d'éclairer cette question d'une lumière nouvelle? En partie, oui, et essentiellement pour les raisons évoquées ci-dessus. Puisque les étendues intraindividuelles sont, en moyenne, pratiquement aussi élevées que les étendues interindividuelles, alors les temps t « choisis » pour mesurer les niveaux de sévérité des examinateurs peuvent très bien avoir un impact important sur les résultats des analyses faites à partir de ces données. Considérons, comme exemple, les graphiques chronologiques des niveaux de sévérité de la figure 4.17. Les examinateurs 1H et 4F ont, au temps 1, un niveau de sévérité respectif de -1,25 et -2,12 logits et l'écart interindividuel est de 0,87 logit. Ces examinateurs ont, au temps 2, des niveaux de sévérité beaucoup plus près l'un de l'autre, respectivement -1,48 et -1,73 logit, pour un écart interindividuel de 0,25 logit. Nous pourrions facilement imaginer qu'une intervention ait lieu entre le temps 1 et le

temps 2 et qu'une réduction de l'écart interindividuel de 0,62 logit soit par la suite attribuée à cette intervention, ce qui ferait l'impasse sur le fait que, au temps 3, les niveaux de sévérité de ces examinateurs retournent pratiquement à leur valeur initiale, respectivement -1,25 et -2,08 logits, pour un écart interindividuel de 0,83 logit. Comme le niveau de sévérité de plusieurs examinateurs de cette thèse fluctue passablement d'un temps de mesure à l'autre en l'absence d'interventions cherchant spécifiquement à modifier ce niveau de sévérité (c.-à-d. une formation), alors il serait, *stricto sensu*, impossible d'attribuer une efficacité quelconque à une éventuelle formation dispensée à ceux-ci, à moins de pouvoir montrer que la modification du niveau de sévérité postérieure à la formation ne perdure sur une longue durée. Il y a bien des examinateurs ayant, pour certaines périodes, un niveau de sévérité extrêmement stable, mais pour savoir cela, encore faut-il mesurer l'évolution temporelle du niveau de sévérité de ces examinateurs. Sans information *a priori* sur l'évolution temporelle du niveau de sévérité d'un examinateur, impossible d'attribuer une efficacité à une intervention ponctuelle censée modifier son niveau de sévérité, puisque ce dernier peut très bien fluctuer en l'absence d'intervention. Cela revient à dire que les résultats de cette thèse appuient le scepticisme exprimé aux sections 2.3.3, 2.3.4 et 2.3.5 quant aux affirmations concernant l'efficacité de telles interventions.

5.2 Descriptions de l'évolution temporelle des niveaux de sévérité

Les données de cette thèse et les 84 séries chronologiques qui en sont tirées mènent à certaines généralisations dignes de mention. Les 3 ensembles de données obtenus à partir des 3 ensembles de données, A, B et L, ont des profils légèrement différents. Les mesures obtenues à l'aide du modèle Rasch à multifacettes donnent lieu à des valeurs assez similaires pour A et L quant à l'étendue des niveaux d'habileté des candidats et de sévérité des examinateurs. Les valeurs, pour B, sont un peu plus restreintes (voir les tableaux 4.1, 4.5 et 4.9), tant pour les niveaux d'habileté des

candidats que pour les niveaux de sévérité des examinateurs. Conséquemment, pour les 84 séries chronologiques du niveau de sévérité étudiées dans cette thèse, les séries des données B ont un écart type inférieur aux séries des données A ou L (moyennes pour A et L = 0,16 logit, moyenne pour B = 0,12 logit). Il n'est malheureusement pas possible de savoir si cette différence dans les notes accordées pour les 3 critères B relève des examinateurs et de leur compréhension de ces 3 critères et de la tâche associée, ou alors de la performance des candidats pour cette tâche ou des caractéristiques formelles de celle-ci, puisque cette thèse est la seule étude, à ce jour, portant sur l'épreuve d'expression orale du TEF. Toutes les études publiées jusqu'ici portent sur l'épreuve d'expression écrite, qui a évidemment des tâches et critères d'évaluation différents, ce qui rend difficile toute comparaison (Artus et Demeuse, 2008 ; Artus, Demeuse, Maréchal, Casanova, Crendal, Desroches et Holle, 2011 ; Casanova et Demeuse, 2011, 2016).

Une autre caractéristique générale à relever est le pourcentage de séries chronologiques à peu près normalement distribuées. Si nous éliminons les 6 séries chronologiques de la modélisation temporelle du 2010-10 au 2013-03, qui n'ont que 10 temps de mesure, alors 57 des 78 (73 %) séries chronologiques restantes sont approximativement normalement distribuées. La répartition de ces 78 séries et du nombre de séries considérées comme normalement distribuées, par modélisation temporelle selon leur ordre à la section « Résultats », est : 29/36 ; 3/6 ; 5/6 ; 11/15 et 9/15. Plus spécifiquement, ces séries sont quasi symétriques, ce qui signifie que, au cours d'une période de temps, ces séries sont aussi souvent au-dessus qu'au-dessous de leur moyenne générale. Bref, à un temps *t* quelconque, un examinateur dont la série est symétrique a autant de chance d'être plus clément ou plus sévère qu'à l'habitude. De même, les valeurs extrêmes sont rares : environ 96 % des valeurs sont situées à 2 écarts types de la moyenne, ce qui correspond presque à la valeur théoriquement attendue pour une distribution parfaitement normale. Qui plus est,

plusieurs valeurs extrêmes sont des artefacts causés par le faible nombre de candidats évalués lors d'un temps *t* (p. ex. section 4.4.1, les 3 séries 1H ou section 4.5.1, série 5F A) et les véritables valeurs extrêmes, inexplicables, sont rares quoique présentes (p. ex. section 4.4.1, série 5F A ou section 4.6.1 série 7F L).

Un autre élément important des séries chronologiques du niveau de sévérité des examinateurs est la faiblesse des tendances linéaires globales de ces séries chronologiques. La figure 5.1 montre la distribution, pour chaque ensemble de données (A, B et L), des 28 valeurs, en logit, représentant la différence de la valeur initiale de la tendance linéaire globale et sa valeur au temps *t*, où *t* égale le nombre de temps de mesure d'une série chronologique donnée.

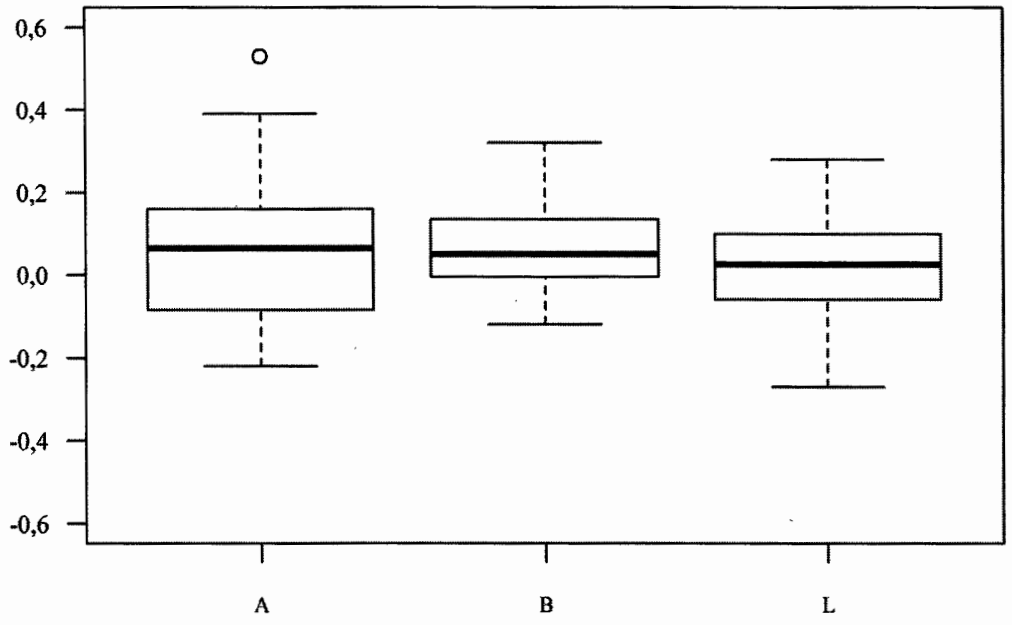


Figure 5.1 : Diagrammes en boîte à moustaches des valeurs, en logit, représentant la différence entre la valeur de la tendance linéaire globale au dernier et au premier temps des 28 séries chronologiques A, B et L.

Les 28 valeurs des séries L sont symétriquement distribuées autour de 0, tandis que les valeurs des séries A et B ont une moyenne et une médiane légèrement supérieures

à 0, ce qui montre qu'il y a un peu plus d'examinateurs dont la série a une tendance linéaire globale positive que négative pour les données A et B. Les valeurs, en logit, de ces tendances linéaires globales sont toutefois très faibles et n'ont, dans la plupart des cas, qu'une importance pratique négligeable. Rappelons qu'une valeur de 0,20, en logit, correspond approximativement à une valeur de 0,5 sur l'échelle d'appréciation de 0 à 20 utilisée par les examinateurs[26]. Une telle différence, sur l'ensemble de la période étudiée, ne représente pas un changement important du niveau de sévérité d'un examinateur, déjà que la capacité d'un examinateur à différencier les 21 échelons possibles est douteuse[27]. Cela signifie que, globalement, les niveaux de sévérité des examinateurs de cette thèse sont temporellement stables, à quelques exceptions près ; certainement, la série 18F A de la modélisation temporelle du nombre de candidats évalués, dont la différence entre le dernier et le premier temps est égale à 0,53 logit, est instable en ce qu'elle montre une augmentation continue du niveau de sévérité. Il y a un total de 5 séries ayant une différence égale ou supérieure à 0,30 logit, incluant la série 18F A susmentionnée. Il n'y a par contre aucune série ayant une différence égale ou inférieure à -0,30 logit. Il y a bien 11 tendances linéaires locales parmi les 84 séries chronologiques de cette thèse, mais c'est somme toute peu eu égard au nombre de séries et de temps de mesure impliqués.

Comment expliquer cette stabilité temporelle globale du niveau de sévérité de ces examinateurs? Encore une fois, la littérature n'offre aucune piste de réponse, puisque toutes les études antérieures comptent de 1 à 8 temps de mesure, ce qui ne permet pas de juger de l'évolution temporelle du niveau de sévérité de leurs examinateurs. La stabilité temporelle des examinateurs de cette thèse pourrait s'expliquer par la manière particulière dont les examinateurs du TEF évaluent l'épreuve d'expression orale. Rappelons qu'il y a toujours 2 examinateurs pour évaluer chaque candidat, que

[26] Pour les valeurs comprises à peu près entre 5 et 17 sur l'échelle d'appréciation.
[27] Nous avons travaillé, durant plusieurs années, comme examinateur avec cette grille et le choix entre deux notes consécutives est parfois très difficile, voire aléatoire.

chaque examinateur note indépendamment chaque performance et que, suite à la notation indépendante, les 2 examinateurs se consultent, procèdent à un arbitrage et notent chacun des 12 critères d'évaluation de manière consensuelle. Cela fait en sorte que les examinateurs développent très rapidement des attentes par rapport à la manière de noter de leurs collègues et que cet arbitrage pourrait influencer le niveau de sévérité des examinateurs en incitant ceux-ci à toujours viser un consensus et à noter de manière « stable ». Cet échange constant de points de vue et de justifications entre les examinateurs s'apparente au processus de modération sociale, où les examinateurs développent leur compréhension et leur maîtrise du référentiel de compétences par l'entremise de la coévaluation des candidats (Laveault et Yerly, 2017). Non seulement les examinateurs développent des attentes communes quant aux notes accordées par leurs collègues, mais ils acquièrent de même une compréhension commune du référentiel de compétences et des échelles de niveaux de compétence. Ces dynamiques sont probablement d'autant plus efficaces que le centre d'où proviennent les données de cette thèse n'a eu, au cours de la période d'où proviennent les données de cette thèse, qu'un maximum de 12 examinateurs actifs, ce qui fait qu'il n'y avait normalement qu'un seul degré de séparation entre n'importe quelle paire d'examinateurs donnée. Cela signifie que chaque examinateur a travaillé avec chaque autre examinateur ou avec un examinateur ayant lui-même travaillé avec tous les autres examinateurs. Plusieurs caractéristiques structurelles convergent pour expliquer l'apparente homogénéité des niveaux de sévérité des examinateurs.

Notons que cette manière de faire, où les examinateurs travaillent ensemble et procèdent systématiquement à un arbitrage, n'est pas commune. La seule étude, parmi les 39 recensées à la section 2.3, qui utilise ce procédé est celle de Meier (2014), et cette recherche porte sur une situation d'évaluation localement développée et utilisée par une université colombienne et elle ne possède qu'un seul temps de mesure. Il est toutefois intéressant de constater que les 4 examinateurs de cette étude

ont un écart interindividuel minuscule de 0,10 logit pour ce temps de mesure. Est-ce que cet écart serait dû à ce procédé? D'autres situations d'évaluation font appel à plus d'un examinateur par performance, mais l'arbitrage, lorsqu'il est nécessaire, est fait par un examinateur supplémentaire, à l'insu des examinateurs initiaux (*TOEFL*, *iELTS*). La dynamique du renforcement du consensus est donc absente.

Un autre facteur qui pourrait expliquer cette relative stabilité temporelle est l'utilisation d'une grille d'évaluation analytique où chaque critère est noté individuellement. Pour chaque examinateur, 3 niveaux de sévérité sont étudiés dans cette thèse, soit les niveaux A, B et L. Ceux-ci sont obtenus à partir des notes respectivement accordées pour les 3 critères d'évaluation concernant la première tâche communicationnelle, pour les 3 critères d'évaluation de la seconde tâche communicationnelle et pour les 6 critères d'évaluation concernant les qualités linguistiques de l'ensemble de la performance du candidat. La modélisation Rasch à multifacettes utilisée pour estimer les niveaux de sévérité fait en sorte que les niveaux sont estimés à partir de la *somme* des notes accordées aux 3 (A et B) ou aux 6 (L) critères d'évaluation. Il est tout à fait possible que, par exemple, un examinateur soit tantôt un peu plus sévère pour le critère 1, un peu plus clément pour le critère 2 et plus sévère pour le critère 3, mais que, d'une journée à l'autre, ces rapports de sévérité changent quelque peu, de manière à ce que, au final, ces différences de sévérité se neutralisent en partie ou en totalité. Une telle conception est compatible avec une définition strictement opérationnelle de ce qu'est la sévérité, définition formulée par Myford et Wolfe (2003) et retenue par cette thèse et par toutes les études recensées qui présentent une définition minimale de ce qu'est la sévérité (voir section 2.1).

Prenons, par exemple, les 6 critères d'évaluation linguistique. Rien n'empêche qu'un examinateur ait des niveaux de sévérité différents, voire très différents, dans

l'évaluation de la prononciation et de la maîtrise lexicale, cet examinateur étant très sévère pour la prononciation et très clément quant à la maîtrise lexicale. Si tel est le cas, alors le niveau de sévérité obtenu à partir de la combinaison des notes accordées à ces 2 critères sera près, en moyenne, de 0, puisque la sévérité du critère 1 et la clémence du critère 2 s'« annuleront ». Cette supposition est étayée par les résultats des recherches ayant étudié les interactions entre les critères d'évaluation et les examinateurs et qui ont constaté plusieurs cas de sévérité différentielle en fonction des critères d'évaluation (Eckes, 2008, 2012 ; Knoch *et al.*, 2015 ; Schaefer, 2008 ; Wigglesworth, 1993). Il y a plusieurs cas, dans ces études, de patrons de sévérité différentielle où un même examinateur est plus clément pour certains critères et plus sévère pour d'autres, c'est donc un phénomène attesté. D'un autre côté, dans ces études, les différences de sévérité relèvent souvent de l'opposition entre les critères d'évaluation communicationnels et linguistiques, mais pas toujours. Certains examinateurs de Wigglesworth ont ainsi des niveaux de sévérité différents pour la phonologie et la grammaire, qui sont pourtant deux critères d'évaluation linguistique, et ces biais vont de 0,04 à 1,17 logit en valeur absolue. Il serait donc possible que ce phénomène affecte certains des examinateurs de cette thèse et qu'il ait un effet perceptible sur les niveaux de sévérité L. Cela aurait pour effet de diminuer la variance des séries L, puisqu'il y aurait « neutralisation » des différents niveaux de sévérité associés à chacun des 6 critères d'évaluation constituant la note L, ce qui ferait en sorte que les valeurs des séries L auraient davantage tendance à être groupées autour de la moyenne. Ce n'est pas ce qui est observé dans cette thèse, comme le montre la figure 5.2, illustrant la distribution des écarts types des 28 séries chronologiques A, B et L. Les écarts types ont été retenus plutôt que les variances, car ces dernières ont des valeurs très faibles (p. ex. 0,02 ou 0,03) rendant difficile toute visualisation.

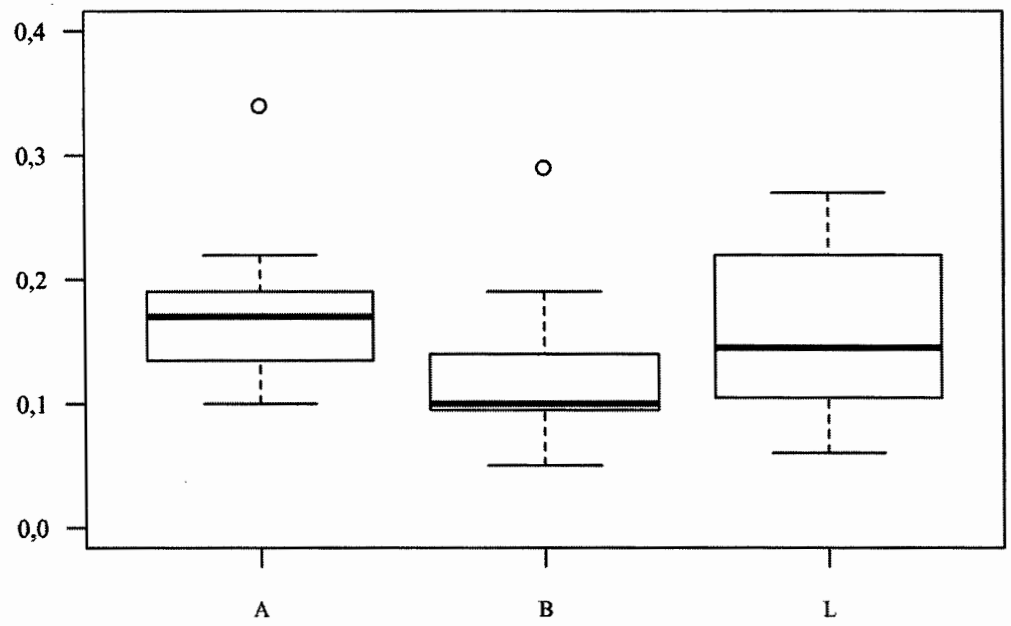


Figure 5.2 : Diagrammes en boîte à moustaches des écarts types, en logit, des 28 séries chronologiques A, B et L.

Les 28 séries L ont une médiane légèrement inférieure à la médiane des 28 séries A, mais leurs moyennes sont pratiquement identiques (0,16 logit arrondi à la 2^{e} décimale) et la distribution de l'ensemble des valeurs est incompatible avec la supposition selon laquelle le nombre supérieur de critères d'évaluation participant au calcul de la note L et l'hétérogénéité de ceux-ci feraient en sorte que la variance des séries L serait inférieure aux variances des séries A et B. Les résultats de cette thèse ne permettent pas d'explorer plus avant cette question ; il faudrait pour cela modéliser l'évolution temporelle du niveau de sévérité estimé pour chaque critère d'évaluation, et non plus à partir de la note combinée de 3 ou 6 critères. Cela serait faisable, mais aurait pour conséquence négative importante de mener à des valeurs estimées du niveau de sévérité ayant des erreurs types très grandes, ce qui diminuerait d'autant l'intérêt de ces analyses. Dans cette thèse, les valeurs de niveau de sévérité sont estimées à partir de la combinaison de 3 ou 6 critères et fait que, pour chaque temps de mesure, la quantité d'information est assez élevée et les erreurs types assez faibles

(0,15 logit ou moins, voir section 3.4.1). Estimer les niveaux de sévérité à partir de chaque critère d'évaluation mènerait à des erreurs types beaucoup plus élevées, ce qui fragiliserait les résultats obtenus (p. ex. les intervalles de confiance à 95 % de chacune des valeurs estimées couvriraient l'ensemble de l'étendue intraindividuelle des niveaux de sévérité.). Nous devons donc nous résoudre à spéculer sur ce possible effet et à laisser à de futures études la tâche d'approfondir cette question.

Les derniers résultats que nous voulons commenter dans cette section concernent les résultats des modélisations AMMI des séries chronologiques de cette thèse. Tel que synthétisé à la section 4.8.2, 30 des 78 séries chronologiques ayant au moins 12 temps de mesure ont un modèle AMMI non nul où au moins l'un des coefficients p, d ou q est d'ordre 1 ou plus. Les 48 autres séries sont modélisées par le modèle nul, correspondant simplement à la moyenne générale de la série à laquelle s'ajoute une erreur à chaque temps de mesure. Pour les modèles non nuls, tous les coefficients obtenus, sauf 3, sont d'ordre 1 ou 2, ce qui correspond à ce que l'on retrouve en général dans les résultats des études utilisant la modélisation AMMI en psychologie (Boswell, Anderson et Barlow, 2014 ; Hamaker, Grasman et Kamphuis, 2016 ; Snippe *et al.*, 2015) ou en sciences sociales en général (Shin, 2017). Ces résultats ne surprennent pas, puisqu'il est plausible que d'éventuels liens entre les niveaux de sévérité se trouvent entre les temps t et $t + 1$ ou $t + 2$, mais beaucoup moins entre t et $t + 3$ ou $t + 4$..., trop éloignés dans le temps. La répartition des 30 modèles non nuls parmi les 5 modélisations temporelles est, elle, plus intrigante. Voici le nombre de séries ayant un modèle non nul pour les 5 modélisations temporelles susmentionnées : 12/36 ; 3/6 ; 3/6 ; 6/15 et 6/15. Cette répartition semble systématique, mais n'est-elle que le fruit du hasard? L'approche novatrice retenue par cette thèse fait que nous n'avons aucun résultat antérieur auquel comparer les nôtres. D'un côté, cette régularité pourrait laisser croire que les modèles obtenus ne sont que des « faux positifs », ce qui expliquerait les pourcentages de modèles non nuls relativement

constants d'une modélisation temporelle à une autre. Cela expliquerait aussi pourquoi certains examinateurs ont des modèles AMMI différents d'une modélisation temporelle à une autre (p. ex. voir les modèles de 5F, présente dans 4 modélisations temporelles différentes, au tableau 4.37). Ces résultats laissent donc croire que les modèles non nuls obtenus ne seraient que le fruit du hasard. En revanche, les différents modèles de certains examinateurs sont remarquablement constants d'une modélisation temporelle à une autre (les modèles de 4F et 13H, présents dans 3 modélisations temporelles). Cette constance serait surprenante si les modèles obtenus n'étaient qu'un résultat aléatoire ; ces résultats, eux, suggèrent plutôt que les modèles non nuls ne sont pas simplement des « faux positifs » et qu'ils révèlent quelque chose à propos de l'évolution temporelle du niveau de sévérité de ces examinateurs. Il n'est toutefois pas possible de trancher ce débat en faveur de l'une de ces deux positions contradictoires, si ce n'est qu'il est fort possible, comme toujours en psychologie et en éducation, que ces contradictions soient dues à des différences individuelles.

5.3 Examinateurs débutants et expérimentés

En accord avec le sens commun et les résultats de nombreuses études montrant un lien positif entre la performance et l'expérience professionnelle en général, quel que soit le domaine (Alessandri, Borgogni et Truxillo, 2015 ; Sturm, 2003), plusieurs des études recensées dans cette thèse ont cherché à savoir s'il y avait des différences entre les examinateurs en fonction de leur expérience professionnelle. Or, toutes ces études n'ont examiné que les différences de niveau de sévérité entre les examinateurs débutants et expérimentés, et ce à un seul temps de mesure (Attali, 2016 ; Hsieh, 2011), deux (Myford *et al.*, 1996) ou trois (H. J. Kim, 2011). Aucune différence significative n'a été trouvée dans ces études quant au niveau de sévérité d'examinateurs débutants et expérimentés. Lim (2009, 2011), lui, a été au-delà de la seule question des différences interindividuelles pour étudier longitudinalement l'évolution du niveau de sévérité d'examinateurs débutants par rapport au niveau de

leurs collègues expérimentés. Globalement, les résultats de ses deux études montrent que certains examinateurs débutants, mais pas tous, ont, lorsqu'ils commencent à travailler, des niveaux de sévérité quelque peu éloignés des niveaux de leurs collègues, mais que ces différences se résorbent assez rapidement. De même, les variances des séries chronologiques des examinateurs débutants et expérimentés sont similaires, si l'on élimine les valeurs extrêmes du premier temps de mesure pour certains examinateurs débutants.

Nos résultats sont en accord avec les résultats de la littérature. Il y a définitivement des examinateurs ayant une période d'ajustement initial au cours de leur insertion professionnelle et ces ajustements peuvent concerner à la fois le niveau de sévérité lui-même, sa volatilité[28] ainsi que sa variance. Ainsi, pour la modélisation temporelle en fonction du nombre de candidats évalués, nous remarquons la présence d'ajustements initiaux du niveau de sévérité pour les séries chronologiques des examinatrices 5F L, 6F L ainsi que 8F A et L. Certaines de ces séries ont aussi une variance plus importante au début de la série, lorsque l'examinatrice est débutante, que lors du reste de la série, après l'ajustement initial (8F A, 17F L et 18F B). D'autres ont également une tendance linéaire locale prononcée (tendance d'un sous-échantillon de temps de mesure), différente de la tendance linéaire globale de la série, ce qui peut être un autre signe de fluctuations liées au manque d'expérience professionnelle (5F A et L, 8F L). Les mêmes résultats s'observent pour les 2 modélisations temporelles chronologiques pour lesquelles il y a 1 (2010-10 au 2013-03) ou 2 examinateurs débutants (2012-12 au 2013-09). Dans le premier cas, l'examinatrice 4F a, pour le niveau de sévérité A, une série chronologique plus volatile que celle de son collègue. Dans le second cas, les séries A des examinateurs débutants (13H et 15F) sont plus volatiles et ont une plus grande variance que celles de leurs collègues expérimentés. En contrepartie, il ne semble pas y avoir de lien

[28] La valeur des différences entre les valeurs consécutives du niveau de sévérité.

systématique entre le fait d'être débutant et la sévérité. Dans ces 2 modélisations temporelles, regroupant 3 examinateurs débutants, donc 9 séries chronologiques, il n'y pas de différences importantes et constantes entre les niveaux de sévérité des examinateurs débutants et expérimentés.

Au final, il semble que la période d'adaptation initiale ne dure pas très longtemps, du moins en ce qui concerne les valeurs extrêmes du niveau de sévérité. Par exemple, les séries chronologiques L des examinatrices 5F, 6F et 8F, ont 1 ou 2 valeurs anormalement élevées lors des 2 premiers temps de mesure. Le niveau de sévérité de ces 3 séries devient « normal » après 3 temps de mesure, représentant donc l'évaluation de seulement 30 candidats. Finalement, nous aimerions contraster le phénomène de l'insertion professionnelle pour un examinateur débutant avec celui de l'insertion professionnelle dans un nouveau milieu de travail pour un examinateur expérimenté. Parmi les 12 examinateurs ayant évalué au moins 100 candidats, 2 ont commencé à travailler au centre de test d'où proviennent les données de cette thèse après avoir travaillé quelques années dans d'autres centres de test, situés en Europe de l'Est. L'une (12F) a commencé à travailler en novembre 2013 et l'autre en janvier 2014 (20F). La première (12 F) a clairement connu un processus d'adaptation pour l'évaluation des critères linguistiques (L). La série chronologique de son niveau de sévérité L a, dans ses 4 premiers temps de mesure, le maximum et le minimum de la série complète. L'écart intraindividuel, si l'on inclut ces deux valeurs extrêmes, est de 1,48 logit. Cet écart diminue à 0,69 logit si l'on exclut ces deux valeurs extrêmes. C'est toutefois le seul cas d'adaptation initiale : les 2 autres séries chronologiques de 12F ne montrent aucun signe d'adaptation initiale, pas plus que les 3 séries de 20F. Il s'agit néanmoins d'un résultat supplémentaire montrant que, dans certains cas, le manque de familiarité avec les pratiques liées à une situation d'évaluation peut mener à un jugement évaluatif déphasé par rapport au jugement évaluatif des autres

examinateurs, que ce manque de familiarité soit dû au fait d'être débutant ou nouveau dans un centre de test donné.

5.4 La modélisation du temps

La comparaison des résultats obtenus à partir des différentes modélisations temporelles de cette thèse mène à un problème fondamental inhérent à toute recherche portant sur l'évolution temporelle du niveau de sévérité : comment modéliser le temps, à quelle durée devrait correspondre chaque temps de mesure d'une modélisation? Dans plusieurs domaines, cette question est de peu d'importance, car une unité temporelle naturelle s'offre à l'analyste. En économétrie, par exemple, il est naturel d'utiliser les données où chaque temps de mesure correspond à une journée, ce qui coïncide avec les heures d'ouverture des marchés financiers d'un fuseau horaire donné. De même, les données géologiques ou hydrologiques utilisent des données saisonnières ou mensuelles correspondant au cycle naturel des dynamiques pertinentes. Même dans les études en psychopathologie, la journée fait office d'unité temporelle logique et largement utilisée par les études utilisant des données longitudinales. Il n'y a pas de telle unité de temps naturelle pour étudier l'évolution temporelle du niveau de sévérité d'examinateurs. Ce qui s'en approche le plus serait d'utiliser la « journée de travail » comme unité temporelle, mais une telle idée rencontre immédiatement deux obstacles majeurs. Cela demanderait que les examinateurs évaluent à peu près le même nombre de candidats d'une journée à l'autre et il faudrait que les journées de travail soient placées à intervalles réguliers (p. ex. chaque mardi), car la modélisation de séries chronologiques à intervalles irréguliers est extrêmement complexe et demande beaucoup de temps de mesure (Belcher, Hampton et Tunnicliffe Wilson, 1994 ; Rehfeld, Marwan, Heitzig et Kurths, 2011). Bref, le choix de l'unité temporelle est difficile et aucune solution idéale ne s'impose.

Il y a également la tension irréductible créée par deux besoins qui s'opposent. Puisque les niveaux de sévérité doivent être estimés (ils ne peuvent être directement mesurés ou dénombrés), il faut le plus de données possible pour avoir des valeurs estimées précises. Par ailleurs, il faut que chaque temps de mesure ne contienne pas un nombre de candidats évalués tel que des dynamiques locales, propres à l'évaluation de quelques candidats, soient écrasées dans le nombre total de candidats évalués pour ce temps de mesure. De surcroît, plus il y a de candidats évalués par temps de mesure, moins il y a de temps de mesure au total. Or, une modélisation précise des séries chronologiques demande le plus grand nombre de temps de mesure possible. Supposons, par exemple, un examinateur très sévère avec une douzaine de candidats consécutifs, mais très clément avec la douzaine suivante. La valeur du niveau de sévérité estimée à partir des ±24 candidats évalués sera à peu près neutre, et les variations propres à chaque douzaine de candidats disparaitront. Un temps de mesure englobant moins de candidats évalués aurait révélé cette dynamique intéressante. Bref, le besoin de précision s'oppose au besoin d'avoir des temps de mesure ne comprenant pas « trop » de candidats évalués. S'ajoutent à ces considérations fondamentales deux problèmes méthodologiques qu'il faut surmonter. Premièrement, la nécessité de modéliser le temps de manière à ce que tous les examinateurs étudiés aient évalué au moins 1 candidat à chaque temps de mesure. Deuxièmement, il faut que tous les examinateurs étudiés dans une modélisation temporelle soient liés par l'entremise des candidats qu'ils ont évalués conjointement, sinon leur niveau de sévérité ne peut pas être estimé sur une même échelle.

Bien que, pour un ensemble de données, il puisse y avoir plusieurs modélisations temporelles répondant à tous les besoins susmentionnés, un problème restera toujours insoluble : que faire des résultats contradictoires provenant des résultats de différentes modélisations temporelles? Puisque, pour un ensemble de données, aucune modélisation temporelle n'a de statut privilégié par rapport aux autres modélisations

temporelles possibles à partir de ces données, comment interpréter des résultats contradictoires? Par exemple, pour cette thèse, les modèles AMMI obtenus pour les séries de l'examinatrice 20F proviennent des mêmes données brutes, seule la modélisation temporelle changeant. Or, ces modèles sont assez différents, mais un modèle n'est pas plus légitime que l'autre. La situation diffère ici du problème classique en statistique de l'indétermination d'un modèle M étant donné les valeurs X, puisque l'on peut toujours ajouter une contrainte pour réduire l'indétermination (p. ex. le principe de parcimonie, qui dicte de choisir, à qualité d'adéquation égale entre deux modèles et les données, le modèle ayant le moins de paramètres). Une telle solution n'est pas possible ici, car le problème est d'une autre nature[29].

Pour cette thèse, cette absence de modélisation temporelle privilégiée mène à des résultats difficiles à réconcilier, particulièrement en ce qui concerne la question de l'étendue intraindividuelle du niveau de sévérité d'un examinateur et du rapport entre cette étendue intraindividuelle et l'étendue interindividuelle de référence. Certains examinateurs ont des rapports intra-/interindividuels très différents selon la modélisation temporelle utilisée, ce qui empêche toute conclusion ferme à ce sujet. Par exemple, l'examinatrice 5F est présente dans 4 modélisations temporelles différentes et, pour les séries chronologiques B, ses rapports intra-/interindividuels vont de 0,53 à 1,96. Difficile en ces circonstances de statuer sur la stabilité temporelle de l'évolution de son niveau de sévérité. Une opposition similaire se trouve, dans nos résultats, pour l'écart type des séries chronologiques du niveau de sévérité L. Les 12 séries chronologiques provenant de la modélisation où chaque intervalle de temps correspond à l'évaluation de 10 candidats ont un écart type médian de 0,19 logit, tandis que les 16 séries chronologiques provenant des 5 modélisations où le temps correspond à une période de temps chronologique fixe ont un écart type médian de

[29] Il serait peut-être possible d'utiliser des méthodes d'analyse plus complexes et raffinées que la modélisation AMMI pour voir si ces différences subsistent, par exemple l'analyse fonctionnelle des données (Ramsay et Silverman, 2005).

0,14 logit. Cela peut sembler peu, mais il est possible que la différence ne soit pas fortuite. Le seul cas théoriquement possible pour lequel le problème du choix de la modélisation temporelle ne se pose pas est celui où la sévérité est conçue comme une caractéristique propre à un individu et qui s'applique uniformément dans toutes les situations d'évaluation, peu importe l'objet évalué ou les critères d'évaluation retenus. C'est la conception traditionnelle épousée par DeCoths (1977), Guilford (1954) et Schriesheim *et al.* (1979) qui est, *de facto*, temporellement invariante. Cela nous amène à la question élémentaire de ce qu'est la sévérité d'un examinateur, question que nous examinons à la section suivante.

5.5 Le concept de « sévérité » des examinateurs

Comme nous l'avons vu dans le contexte théorique, les études recensées sont laconiques quant à la nature de la sévérité. Soit elles ne définissent pas le concept, soit elles le définissent d'une manière purement opérationnelle, à l'instar de Saal *et al.* (1980) ou Myford et Wolfe (2003). Les études de Bachman *et al.* (1995), Bonk et Ockey (2003), ou Casanova et Demeuse (2011) sont des exemples du premier cas, tandis que les études de Cai (2012), Eszter (2007) ou Lim (2009) illustrent le second. Notre thèse a également adopté la définition opérationnelle de Myford et Wolfe. Ce choix est toutefois vulnérable aux critiques classiques faites envers l'opérationnalisme (Bickhard, 2001 ; Feest, 2005 ; Maul et McGrane, 2017). Premièrement, cette définition ne nous renseigne en rien sur ce qu'est la sévérité et, deuxièmement, les résultats obtenus sont tributaires du modèle de mesure utilisé pour opérationnaliser la sévérité. Ainsi, nos résultats ne valent qu'en vertu du modèle de Rasch à multifacettes et l'utilisation d'un modèle différent, par exemple celui proposé par Casanova et Demeuse (2016) pourrait mener à des conclusions différentes – dans une certaine limite, puisque Casanova et Demeuse ont obtenu des corrélations élevées ($r \geq 0{,}91$) entre les valeurs estimées des niveaux de sévérité des examinateurs selon le modèle de Rasch à multifacettes et selon le modèle qu'ils ont proposé. Nos résultats

permettent toutefois de faire quelques inférences quant aux caractéristiques de ce qu'est la sévérité d'un examinateur.

Il est clair que nos résultats sont incompatibles avec une conception unifiée de la sévérité vue comme caractéristique de l'examinateur, stable peu importe l'objet ou le critère d'évaluation. Les résultats des modélisations AMMI et des corrélations croisées intraindividuelles entre les 3 niveaux de sévérité de chaque examinateur contiennent trop de contre-exemples incompatibles avec une telle conception unifiée de ce qu'est la sévérité. Si une telle conception était vraie, alors nous nous attendrions à retrouver, pour un même examinateur, des modèles AMMI similaires pour chacun des niveaux de sévérité (A, B et L), *peu importe* la modélisation temporelle retenue. De même, cette conception mènerait à des corrélations croisées intraindividuelles ayant, pour chaque examinateur, des valeurs similaires, puisque chaque niveau de sévérité (A, B ou L) ne serait que l'instanciation d'un niveau de sévérité général, propre à l'examinateur et constant. Or, nous avons déjà vu que les modèles AMMI obtenus pour un même examinateur différaient substantiellement, dans certains cas, d'une modélisation temporelle à une autre. Quant aux corrélations intraindividuelles entre les niveaux de sévérité, il y a plusieurs examinateurs pour lesquels ces corrélations changent passablement d'une modélisation temporelle à une autre, ce qui est incompatible avec cette conception. Par exemple, les corrélations entre les niveaux A et L de l'examinatrice 5F, au délai 0, pour les 4 modélisations temporelles dans lesquelles elle apparaît, sont de 0,16 ; 0,40 ; 0,54 et 0,01. Le même phénomène se produit pour les corrélations, au délai 0, entre les niveaux de sévérité A et B de l'examinateur 1H, pour les 3 modélisations temporelles où il se trouve : 0,19 ; 0,61 et 0,59. Dans ces deux exemples, des écarts de 0,4 ou 0,5 sont observés, ce qui semble trop important pour être expliqué par l'erreur de mesure ou autre. Considérant que la littérature semble avoir abandonné cette idée d'une sévérité « unique » à chaque personne, il semble que, à l'aune de nos résultats, cette conception puisse être rejetée.

Une autre conception possible de ce qu'est la sévérité est que chaque personne a plusieurs niveaux de sévérité différents, propres à chaque objet d'évaluation. Par exemple, un examinateur aurait un niveau de sévérité *X* pour évaluer les habiletés argumentatives et un niveau de sévérité *Y* dans l'évaluation des habiletés à poser des questions et à obtenir des informations. Chaque examinateur aurait ainsi autant de niveaux de sévérité qu'il y a d'objets évalués possibles et on peut supposer qu'une proximité entre les objets évalués mènerait à une similarité entre les niveaux de sévérité reliés. Les résultats de cette thèse permettent d'étudier indirectement la plausibilité de cette conception, principalement en comparant les corrélations croisées intraindividuelles entre les niveaux de sévérité A et B par rapport aux corrélations entre A et L et B et L. En vertu de cette conception, la plus grande proximité conceptuelle entre les objets évalués en A et B, qui sont différentes habiletés communicationnelles, devrait faire en sorte que les niveaux de sévérité A et B sont plus fortement corrélés que ne le sont les niveaux de sévérité A et L ou B et L. La figure 5.3 synthétise les résultats de cette thèse sur cette question.

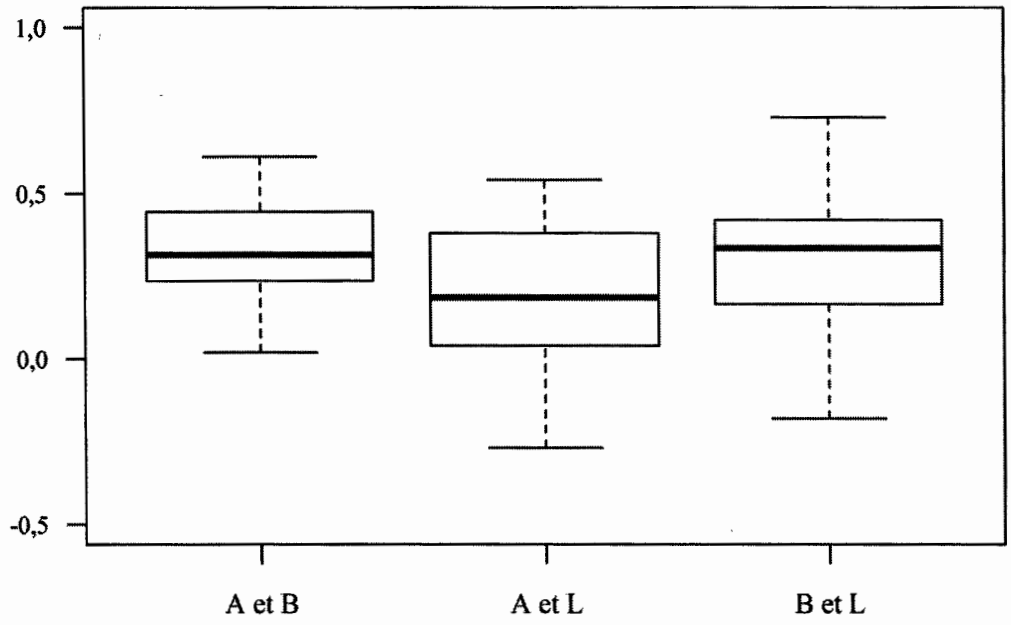


Figure 5.3 : Diagrammes en boîte à moustaches de la distribution des 28 coefficients de corrélations croisées intraindividuelles pour les 3 paires de séries chronologiques A, B et L.

Il n'y a pas de différences importantes entre les coefficients de corrélation A/B et B/L : les médianes sont très proches, de même que les 1er et 3^{e} quartiles. La différence est plus marquée entre les corrélations A/B et A/L, mais cela ne suffit pas pour affirmer que les corrélations entre les séries chronologiques A et B sont supérieures aux corrélations entre les 2 autres paires possibles de séries chronologiques *à cause* de la similarité entre les objets évalués en A et B. Par ailleurs, la très grande majorité des coefficients ont une valeur positive, ce qui montre que les 3 niveaux de sévérité d'un même examinateur sont, généralement, positivement corrélés. Le sont-ils parce que les 3 niveaux de sévérité concernent des objets d'évaluation qui, tous trois, relèvent de l'évaluation en français, langue étrangère? Ou est-ce parce que les personnes tendent à avoir des niveaux de sévérité qui, peu importe l'objet d'évaluation, sont positivement corrélés? Nous ne pouvons que spéculer, les résultats de cette thèse ne nous permettant pas de répondre, même très partiellement, à ces questions.

Malheureusement, aucune étude n'a cherché à étudier le niveau de sévérité d'examinateurs dans différentes situations d'évaluation, chacune ayant des objets d'évaluation différents. Il se pourrait également que les niveaux de sévérité des examinateurs soient liés non pas aux objets d'évaluation, mais bien aux critères d'évaluation. Ainsi, une personne aurait des niveaux de sévérité similaires dans l'évaluation de la grammaire, que ce soit en langue d'enseignement, à l'écrit, qu'en langue étrangère, à l'oral et ce sans égard à la nature discursive du texte évalué. Ces deux suppositions sont compatibles avec nos résultats selon lesquels les niveaux de sévérité sont plutôt temporellement stables, comme en font foi les séries chronologiques étudiées dans cette thèse. La faiblesse relative des tendances linéaires globales, en particulier, est compatible avec l'idée de niveaux de sévérité comme caractéristiques relativement stables d'un examinateur, que ces niveaux de sévérité portent sur un objet ou un critère d'évaluation. Mais ces niveaux de sévérité varient d'un temps à l'autre, peut-être à cause de tous les facteurs hypothétiques suggérés par la littérature, par exemple la fatigue de l'examinateur, les interactions entre les différents éléments de la situation d'évaluation, les divers effets de l'examinateur (halo, séquence, tendance centrale...). La multitude de facteurs impliqués dans chaque évaluation, sur une période de temps *t*, expliquerait ainsi pourquoi 73 % des séries chronologiques sont approximativement normalement distribuées et pourquoi les plateaux sont rares dans les séries chronologiques de cette thèse. Le nombre important de facteurs ayant un impact sur le niveau de sévérité d'un examinateur à un temps donné serait également compatible avec nos résultats contradictoires concernant les corrélations croisées interindividuelles entre les séries chronologiques des niveaux de sévérité d'examinateurs travaillant conjointement durant une période de temps. Cela serait également compatible avec la rareté des valeurs extrêmes des niveaux de sévérité, soit les valeurs situées à plus de 2 écarts types de la moyenne générale. La sévérité d'un examinateur serait donc liée à un objet ou à un critère d'évaluation et serait constituée d'un noyau stable, auquel se grefferait, selon le contexte, l'influence de divers facteurs.

Une telle conception de la sévérité serait aussi cohérente avec les résultats de la littérature portant sur l'efficacité de la formation des examinateurs. Comme nous l'avons vu, la formation donne des résultats mitigés, positifs dans 2 études sur 6 et neutres ou négatifs dans les 4 autres études. Dans ces 6 études, les examinateurs ayant recentré leur niveau de sévérité avaient un niveau très sévère ou clément au temps 1 et ce niveau s'est rapproché du niveau de leurs collègues au temps 2, suite à une formation. Le succès limité de la formation, circonscrit aux seuls niveaux de sévérité extrêmes, pourrait être dû au fait qu'il est facile pour un examinateur se sachant trop sévère de systématiquement hausser les notes qu'il accorde, ce qui règlerait par la force brute son problème de sévérité. Une telle modification de comportement écraserait également l'impact de tous les facteurs potentiels susmentionnés et ferait en sorte que la formation serait vue comme un succès. En revanche, on peut supposer que les examinateurs légèrement trop sévères ou cléments ne pourraient pas adopter un tel comportement par crainte de succomber à l'autre extrême, ce qui ferait en sorte que les facteurs potentiels continueraient à affecter le niveau de sévérité de ces examinateurs, ce qui aurait pour résultat que leur niveau de sévérité pourrait rester stable, augmenter en sévérité ou en clémence. Nous avons vu de tels résultats dans la littérature recensée. Cette conception serait aussi compatible avec les résultats des études de Eckes (2008, 2012) établissant une typologie des examinateurs en fonction des critères d'évaluation auxquels ils accordent plus d'importance. Ces études ont montré que ces examinateurs tendent à avoir des niveaux de sévérité en lien avec les critères d'évaluation qu'ils considèrent importants. Mais nous insistons sur le caractère spéculatif de cette conceptualisation, qu'il faudrait étudier directement. Les analyses et résultats de cette thèse ne permettent qu'un regard oblique, indirect, sur cette question néanmoins fondamentale et au cœur des problèmes pratiques liés aux différences de niveaux de sévérité d'examinateurs.

CHAPITRE VI

CONCLUSION

Cette thèse visait à répondre à la question générale de recherche : « Comment évolue longitudinalement le niveau de sévérité d'examinateurs? ». Comme l'a montré la recension des écrits, plusieurs études laissent croire que le niveau de sévérité d'examinateurs est loin d'être temporellement stable, mais aucune étude antérieure à cette thèse n'a véritablement étudié l'évolution longitudinale du niveau de sévérité d'examinateurs avec un nombre de temps de mesure assez élevé, toutes ces études ayant de 3 à 12 temps de mesure. Or, mieux connaître l'évolution temporelle du niveau de sévérité est important, car la stabilité relative du niveau de sévérité des examinateurs est une condition nécessaire, mais non suffisante, à l'efficacité de leur formation et à la certification de leur compétence, du moins en ce qui concerne leur niveau de sévérité. Face à la paucité de résultats pertinents, cette thèse a retenu 3 objectifs spécifiques de recherche, soit : « Modéliser l'évolution du niveau de sévérité des examinateurs en fonction du nombre de candidats évalués et du temps chronologique », « Comparer l'évolution du niveau de sévérité des examinateurs débutants et expérimentés » et « Comparer l'évolution du niveau de sévérité d'examinateurs travaillant ensemble ».

Les données proviennent de l'épreuve d'expression orale du TEF, un test de français langue étrangère reconnu. Ces données secondaires viennent d'un centre de test canadien et ont été collectées d'octobre 2010 à avril 2014. Un total de 3 333 candidats ont fait l'épreuve d'expression orale et ils ont été évalués par un total de 20

examinateurs, dont 12 ont évalué un minimum de 100 candidats. Chaque candidat a été évalué sur 2 tâches communicationnelles, par 2 examinateurs, à l'aide d'une échelle d'appréciation à 12 critères et 21 échelons. Les 12 critères se regroupent en 3 notes distinctes, obtenues par la sommation des notes accordées à chaque critère : la note A pour la première tâche (3 critères), la note B pour la seconde tâche (3 critères) et la note L pour les qualités linguistiques de l'ensemble de la performance du candidat (6 critères). Les données brutes ont été analysées avec le modèle de Rasch à multifacettes pour obtenir les valeurs estimées du niveau de sévérité des examinateurs selon 6 modélisations temporelles distinctes. Une première où chaque temps de mesure correspond à l'évaluation d'exactement 10 candidats consécutifs et 5 modélisations temporelles chronologiques où chaque temps de mesure correspond à un certain nombre de mois (de ⅓ à 3 mois selon la modélisation). Les valeurs estimées du niveau de sévérité, en logit, ont ensuite été utilisées comme variable dépendante pour la modélisation des séries chronologiques à l'aide de la modélisation AMMI.

Les résultats montrent d'abord que les séries chronologiques du niveau de sévérité tendent à être normalement distribuées ; elles ont une faible asymétrie et les valeurs extrêmes, situées à plus de 2 écarts types de la moyenne, sont rares. Environ un tiers des examinateurs ont des problèmes de dérive temporelle, c'est-à-dire que leur niveau de sévérité est temporellement instable. L'étendue intraindividuelle du niveau de sévérité est, pour les données A, généralement inférieure à l'étendue interindividuelle, mais elle est, pour les données B et L, égale ou supérieure à l'étendue interindividuelle pour près de la moitié des examinateurs. Les examinateurs débutants ont, dans quelques cas, des problèmes d'adaptation initiale et une tendance à avoir les valeurs minimales ou maximales de leur niveau de sévérité au début de leur carrière, mais cela n'est pas systématique : plusieurs examinateurs débutants ne diffèrent en rien, même au tout début de leur carrière, de leurs collègues plus expérimentés. Les

résultats concernant l'évolution du niveau de sévérité d'examinateurs travaillant ensemble au cours d'une période donnée mènent à des conclusions semblables. Certaines paires d'examinateurs ont une corrélation entre leur niveau de sévérité d'une force faible ou moyenne, mais d'autres paires ont une corrélation essentiellement nulle. Bref, pour tous les résultats, il y a des différences importantes entre les examinateurs, ce qui fait toute la richesse d'une telle étude longitudinale.

Plusieurs limites circonscrivent toutefois les résultats de cette thèse. Premièrement, ces résultats ne sont pas généralisables aux autres examinateurs d'examens à forts enjeux, et ce pour plusieurs raisons. Les éléments des situations d'évaluation – type d'épreuve d'expression orale, critères d'évaluation, nombre d'échelons de l'échelle d'appréciation, nombre d'examinateurs évaluant chaque performance – changent d'un examen à un autre et ces éléments peuvent affecter le comportement des examinateurs. La dynamique de la double évaluation à l'aveugle des candidats, suivie d'un arbitrage consensuel ainsi que la formation suivie par les examinateurs étudiés dans cette thèse, tout cela fait en sorte que la « culture évaluative » de ces examinateurs est probablement unique et empêche toute généralisation à d'autres examinateurs, pour ne rien dire de l'échantillonnage non aléatoire de ceux-ci.

Deuxièmement, les données de cette thèse sont des données secondaires. La collecte des données n'a donc pas été planifiée et organisée dans le but d'atteindre les objectifs spécifiques de la thèse, ce qui nuit certainement à la qualité des résultats obtenus. Cela est particulièrement vrai pour le 2^{e} objectif spécifique, cherchant à comparer l'évolution du niveau de sévérité des examinateurs débutants et expérimentés, car les 5 modélisations temporelles en fonction du temps chronologique retenues n'ont que 3 examinateurs débutants dont l'évolution du niveau de sévérité peut être comparée à celle de leurs collègues expérimentés. La modélisation temporelle en fonction du nombre de candidats évalués permet bien de

comparer l'évolution du niveau de sévérité d'un examinateur lorsqu'il est débutant avec l'évolution de son niveau de sévérité après qu'il est devenu expérimenté, mais c'est un pis-aller, car il s'agit d'une comparaison intraindividuelle. Mais le fait que les données soient des données secondaires ne nuit pas qu'à l'atteinte du 2^{e} objectif spécifique de recherche. Les 1er et 3^{e} objectifs spécifiques doivent également composer avec les limites imposées par les données, surtout par rapport au nombre de candidats évalués par chaque examinateur à chacun des temps de mesure d'une modélisation temporelle. Ainsi, pour les 5 modélisations temporelles en fonction du temps chronologique, il arrive qu'un examinateur n'ait évalué qu'un seul candidat à un temps de mesure donné, ce qui mène à une valeur extrême du niveau de sévérité pouvant affecter les analyses subséquentes (modélisation AMMI ou corrélations croisées). Un autre problème découlant de la nature secondaire des données est que cela limite également le nombre de temps de mesure ou d'examinateurs d'une modélisation temporelle. Parmi les 5 modélisations temporelles chronologiques de la thèse, 3 ont moins de 20 temps de mesure et les 2 modélisations ayant plus de 30 temps de mesure n'ont que 2 examinateurs chacune. Il aurait été très intéressant d'avoir 4 ou 5 examinateurs au sein d'une modélisation temporelle ayant 30 ou 40 temps de mesure, mais les données disponibles ne permettaient pas cela. Finalement, le nombre relativement faible de temps de mesure pour certaines modélisations temporelles rend hasardeuse l'interprétation des coefficients de corrélations croisées obtenus pour ces modélisations, la littérature recommandant de 200 à 250 paires de données pour obtenir des coefficients de corrélation stables (Schönbrodt et Perugini, 2013).

L'ensemble de ces limites mène naturellement à nos recommandations quant aux pistes de recherche à explorer. Il faudrait d'abord plusieurs études explorant l'évolution longitudinale du niveau de sévérité d'examinateurs, avec un minimum de 30 temps de mesure, préférablement davantage, afin de constituer une littérature

quelque peu étoffée sur la cette question importante. Cela permettrait une réelle comparaison des résultats des diverses études, ce qui enrichirait notre compréhension de ce phénomène. Il serait également très instructif d'étudier l'évolution temporelle du niveau de sévérité d'examinateurs suivant régulièrement des formations continues visant à réguler leur niveau de sévérité, et ce, afin de voir la durée d'éventuels effets bénéfiques de la formation. Comme nous l'avons vu dans la revue de littérature, la question de l'efficacité de la formation est cruciale et l'état actuel des connaissances ne permet pas de statuer sur celle-ci. Des études longitudinales permettraient certainement de consolider nos connaissances à ce sujet. Dans un même ordre d'idées, des études longitudinales avec davantage d'examinateurs débutants que n'en compte cette thèse permettraient de voir jusqu'à quel point les problèmes d'adaptation initiale sont répandus.

Une autre piste à explorer serait la réplication de cette thèse, mais avec des données provenant d'examinateurs de l'épreuve d'expression écrite du TEF. Cela permettrait de comparer l'évolution longitudinale du niveau de sévérité d'examinateurs partageant plusieurs éléments de la situation d'évaluation, mais différant sur un point important : la notation de l'épreuve d'expression écrite est faite par 2 examinateurs, mais elle est individuelle et il n'y a pas d'arbitrage consensuel fait par les 2 examinateurs eux-mêmes, comme c'est le cas pour l'épreuve d'expression orale (Casanova et Demeuse, 2016). L'arbitrage est plutôt fait par un examinateur tiers, ce qui fait que la dynamique particulière à l'œuvre pour les examinateurs de l'expression orale n'existe pas pour les examinateurs de l'expression écrite. Les résultats d'une telle étude permettraient de vérifier l'hypothèse que nous avons émise quant au rôle potentiel de cette dynamique sur l'homogénéité des niveaux de sévérité des examinateurs de cette thèse. Une autre idée à approfondir serait celle concernant l'étude de l'évolution temporelle du niveau de sévérité relié à chaque critère d'évaluation analytique et non plus à une note composite, comme c'est le cas dans

cette thèse et dans toutes les études antérieures. Les résultats d'une recherche de la sorte apporteraient d'importantes informations indirectes quant au statut du concept de sévérité et aux liens entre objet d'évaluation, critère d'évaluation et niveau de sévérité.

Finalement, il pourrait être intéressant d'explorer l'utilisation de modélisations différentes pour estimer le niveau de sévérité des examinateurs, par exemple en utilisant un modèle linéaire généralisé (De Boeck et Wilson, 2004), un modèle de test à trait latent (Fischer, 1983) ou en ayant recours à un modèle à transitions latentes, une variation des modèles à classes latentes (Collins et Lanza, 2010). Un tel modèle permettrait de représenter directement par un paramètre la probabilité qu'un examinateur aurait de passer d'un niveau de sévérité à un autre, ce qui pourrait apporter des informations utiles sur le phénomène de dérive temporelle du niveau de sévérité.

Ces nombreuses limites montrent bien que, somme toute, cette thèse n'est qu'un modeste premier pas sur la route des études longitudinales du niveau de sévérité d'examinateurs, route amorcée par Lim (2009, 2011). Il est important que cette route ne soit pas un cul-de-sac et que d'autres études poursuivent la quête.

ANNEXE A : TESTS DIAGNOSTIQUES DES SÉRIES CHRONOLOGIQUES

Valeurs *p* des 6 tests diagnostiques utilisés pour vérifier le respect des conditions d'utilisation de la méthode Box-Jenkins. Les colonnes ont les significations suivantes :

TLG : Test de blancheur de Teräsvirta, Lin et Granger. H_0 = il n'y a pas de relation non linéaire dans la série chronologique.

Llung-Box : Test de Llung-Box. H_0 = Les données de la série chronologique sont indépendantes.

BDS : Test de Brock, Dechert et Scheinkman. H_0 = Les données de la série chronologique sont indépendantes et identiquement distribuées.

KPSS : Test de Kwiatkowski, Phillips, Schmidt et Shin. H_0 = La série chronologique est faiblement stationnaire.

ADF : Test augmenté de Dickey et Fuller. H_A = La série chronologique est faiblement stationnaire. C'est le seul des six tests où l'on veut rejeter H_0 et accepter H_A.

B de Bartlett : Test de Bartlett. H_0 = La série chronologique est du bruit blanc.

Les valeurs *p* indiquant une possible violation des conditions d'utilisation de la méthode Box-Jenkins sont en italique, en caractère gras.

Modélisation temporelle en fonction du nombre de candidats évalués

Séries		TLG	Llung-Box	BDS	KPSS	ADF	B Bartlett
	A	0,31	0,55	0,47	0,10	0,01	0,32
1H	B	0,34	0,98	***0,01***	0,10	0,02	0,99
	L	0,67	0,45	0,49	0,10	0,01	0,56
	A	0,35	0,58	0,93	0,10	0,01	0,98
4F	B	0,80	0,48	0,19	0,09	0,01	0,68
	L	0,36	0,94	0,17	0,10	0,01	0,39
	A	0,37	0,89	0,09	0,10	0,01	0,98
5F	B	0,27	0,56	0,96	0,10	0,01	0,80
	L	0,40	0,26	0,16	0,10	0,01	0,44
	A	0,76	0,99	0,69	0,10	0,01	0,97
6F	B	0,60	0,43	***< 0,01***	0,10	0,01	0,99
	L	0,38	0,21	***< 0,01***	0,10	0,01	0,71
	A	0,48	0,87	0,87	0,10	0,01	0,99
7F	B	0,85	0,69	0,16	0,10	0,01	0,50
	L	0,75	0,30	0,87	0,10	0,05	0,65
	A	0,47	0,50	***< 0,01***	0,10	0,01	0,48
8F	B	0,95	0,81	***< 0,01***	0,09	0,01	0,99
	L	0,09	0,73	0,36	0,10	0,01	0,30
	A	0,95	0,93	0,70	0,10	0,01	0,99
12F	B	0,16	0,34	0,79	0,10	0,01	0,96
	L	0,66	0,96	0,45	***0,03***	0,01	0,99
	A	0,26	0,14	0,09	0,10	0,01	0,72
13H	B	0,29	0,25	0,29	0,10	0,01	0,92
	L	0,37	0,66	***< 0,01***	0,10	0,01	0,50
	A	0,10	0,93	***0,01***	0,10	0,01	0,98
15F	B	0,26	0,33	***< 0,01***	0,10	0,01	0,81
	L	0,86	0,17	0,46	0,10	0,01	0,91
	A	0,48	0,65	0,92	0,10	0,01	0,94
17F	B	0,15	0,95	0,06	0,10	0,01	0,85
	L	0,01	0,83	***< 0,01***	0,10	0,03	0,99
	A	0,20	0,40	0,57	0,10	0,01	0,34
18F	B	0,36	0,76	***0,01***	0,10	0,01	0,99
	L	***0,05***	0,97	***0,01***	0,10	***0,06***	1,00
	A	0,63	0,66	***< 0,01***	0,10	0,01	0,87
20F	B	0,41	0,71	0,55	0,10	0,04	0,92
	L	0,96	0,64	***< 0,01***	0,10	0,03	0,36

Modélisation temporelle du 2010-10 au 2013-03

Séries		TLG	Llung-Box	BDS	KPSS	ADF	B Bartlett
	A	***0,03***	0,36	***< 0,01***	0,10	***0,12***	0,99
1H	B	0,19	0,66	***< 0,01***	0,10	0,01	0,61
	L	0,33	***0,05***	***< 0,01***	0,10	0,02	0,34
	A	0,23	0,51	***< 0,01***	0,10	***0,07***	0,91
4F	B	***0,03***	0,94	0,05	0,10	0,01	0,85
	L	***< 0,01***	0,72	***< 0,01***	0,08	0,01	0,83

Modélisation temporelle du 2011-09 au 2013-02

Séries		TLG	Llung-Box	BDS	KPSS	ADF	B Bartlett
	A	0,68	0,41	***< 0,01***	0,10	0,03	0,73
1H	B	***0,03***	0,97	0,51	0,10	0,02	0,99
	L	***0,02***	0,79	0,11	0,10	0,01	0,98
	A	0,57	0,73	0,07	0,10	0,01	0,40
5F	B	0,17	0,97	0,19	0,10	0,01	0,88
	L	0,23	0,51	0,26	0,10	0,01	0,52

Modélisation temporelle du 2012-06 au 2013-11

Séries		TLG	Llung-Box	BDS	KPSS	ADF	B Bartlett
	A	0,90	0,45	0,25	0,10	0,01	0,99
4F	B	0,89	0,94	0,09	0,10	0,01	1,00
	L	0,16	0,47	0,29	0,10	0,01	0,79
	A	0,93	0,83	0,55	0,10	0,01	0,77
5F	B	0,25	0,92	***0,05***	0,10	0,01	0,80
	L	0,12	0,97	0,06	0,10	0,01	0,99

Modélisation temporelle du 2012-12 au 2013-09

Séries		TLG	Llung-Box	BDS	KPSS	ADF	B Bartlett
	A	0,50	0,62	***< 0,01***	0,08	0,05	0,45
4F	B	0,54	0,69	***< 0,01***	0,08	***0,16***	0,98
	L	***0,04***	***0,05***	***< 0,01***	0,10	0,05	0,40
	A	0,12	0,74	0,009	0,10	0,01	0,90
5F	B	0,17	0,42	0,58	0,10	0,01	0,84
	L	***< 0,01***	0,56	0,07	0,10	0,01	0,78
	A	0,36	0,84	0,94	0,10	0,02	0,71
7F	B	0,46	0,78	***< 0,01***	0,10	***0,08***	0,98
	L	0,06	0,56	***< 0,01***	0,10	0,01	0,62
	A	0,90	0,61	***< 0,01***	0,10	0,01	0,99
12F	B	0,20	0,97	0,08	0,10	***0,18***	0,98
	L	0,51	0,80	0,96	0,10	0,01	0,85
	A	0,26	0,28	0,46	0,10	0,01	0,45
13H	B	0,99	0,95	0,03	0,10	0,01	0,94
	L	0,54	0,35	***< 0,01***	0,10	0,01	0,95

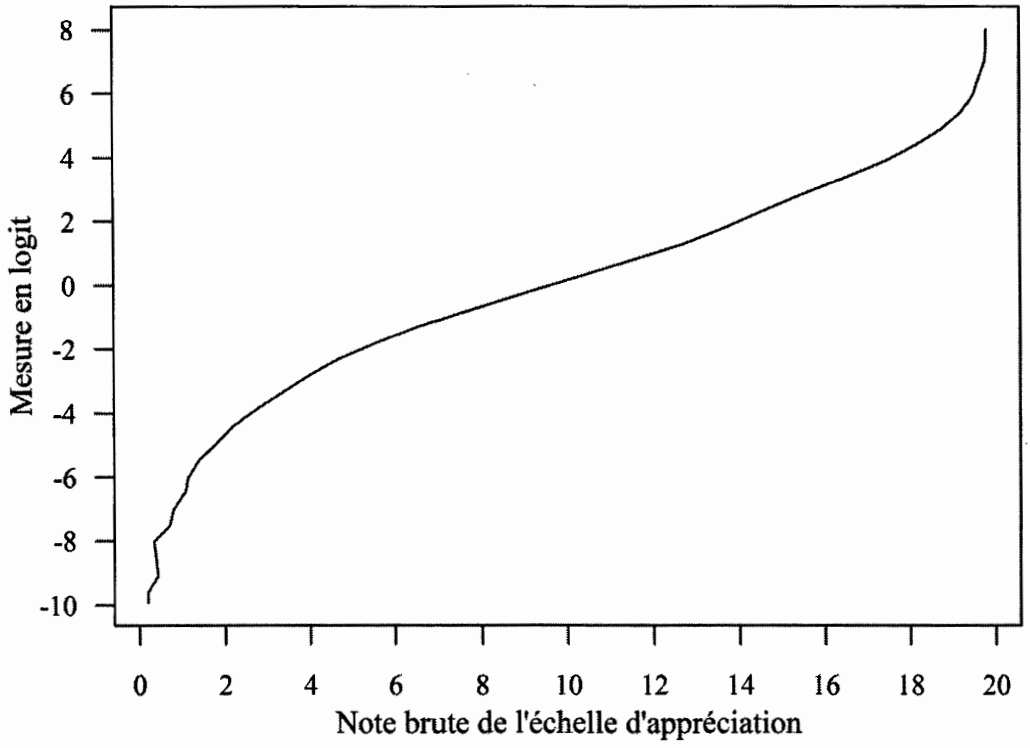
Modélisation temporelle du 2013-11 au 2014-04

Séries		TLG	Llung-Box	BDS	KPSS	ADF	B Bartlett
	A	0,27	0,23	1	0,10	0,01	0,40
13H	B	0,66	0,67	0,18	0,10	0,01	0,96
	L	0,81	0,32	0,42	0,10	0,03	0,96
	A	0,84	0,58	***< 0,01***	0,10	0,01	0,57
15F	B	0,17	0,76	0,93	0,07	0,01	0,64
	L	0,92	0,91	***< 0,01***	0,10	***0,06***	1,00
	A	0,19	0,49	***< 0,01***	0,10	0,01	0,99
17F	B	0,35	0,59	***< 0,01***	0,10	0,01	0,81
	L	***0,03***	0,64	0,52	0,10	0,01	0,99
	A	***< 0,01***	0,91	***< 0,01***	0,10	0,01	0,59
18F	B	0,63	0,99	0,73	0,10	0,04	0,95
	L	0,73	0,80	0,07	0,10	0,01	0,85
	A	0,89	0,82	0,21	0,10	0,02	0,79
20F	B	0,84	0,68	0,99	0,10	0,02	0,96
	L	0,65	0,94	***< 0,01***	0,10	0,01	0,96

ANNEXE B : RELATION ENTRE LES NOTES BRUTES ET LES MESURES EN LOGIT

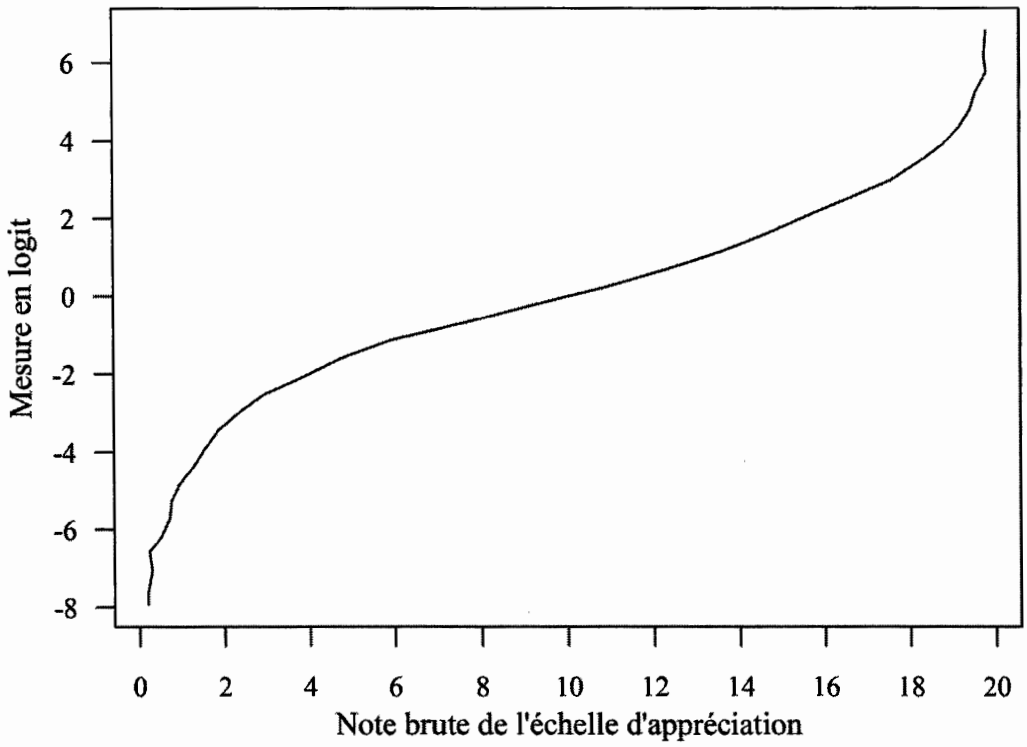
Cette annexe présente les courbes empiriques de la relation entre les valeurs moyennes des mesures en logit et des notes brutes de l'échelle d'appréciation de 0 à 20, et ce pour les 3 ensembles de données (A, B et L).

B.1 : Pour les notes A



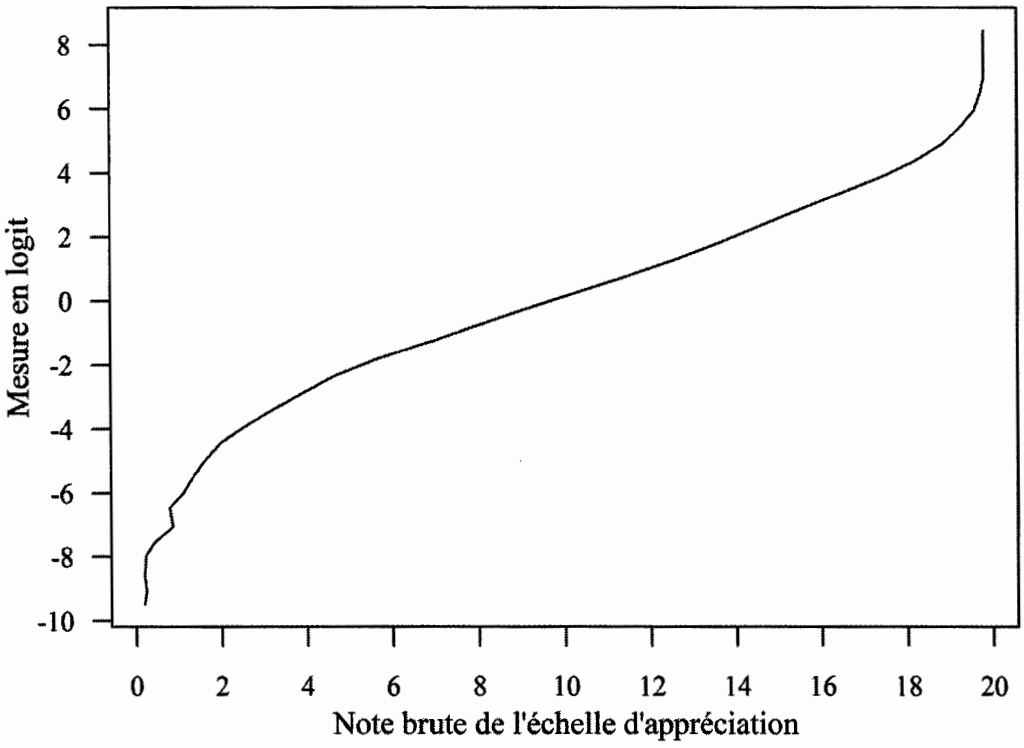
Pour les notes supérieures à 4 et inférieures à 18, la corrélation linéaire est de 0,99. La régression de la note brute sur la mesure en logit a un coefficient de régression de 2,11 (erreur type = 0,04) et le coefficient de détermination ajusté est de 0,99. La relation entre la note brute et la mesure en logit, entre l'étendue de 5 à 17 sur l'échelle d'appréciation peut donc être considérée comme presque parfaitement linéaire.

B.2 : Pour les notes B



Pour les notes supérieures à 4 et inférieures à 18, la corrélation linéaire est de 0,99. La régression de la note brute sur la mesure en logit a un coefficient de régression de 2,87 (erreur type = 0,10) et le coefficient de détermination ajusté est de 0,99. La relation entre la note brute et la mesure en logit, entre l'étendue de 5 à 17 sur l'échelle d'appréciation peut donc être considérée comme presque parfaitement linéaire.

B.3 : Pour les notes L



Pour les notes supérieures à 4 et inférieures à 18, la corrélation linéaire est de 0,99. La régression de la note brute sur la mesure en logit a un coefficient de régression de 2,08 (erreur type = 0,03) et le coefficient de détermination ajusté est de 0,99. La relation entre la note brute et la mesure en logit, entre l'étendue de 5 à 17 sur l'échelle d'appréciation peut donc être considérée comme presque parfaitement linéaire.

RÉFÉRENCES

Alessandri, G., Borgogni, L. et Truxillo, D. M. (2015). Tracking job performance trajectories over time: A six-year longitudinal study. *European Journal of Work and Organizational Psychology*, *24*(4), 560-577.

American Educational Research Association, American Psychological Association et National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC : American Educational Research Association.

American Psychological Association, American Educational Research Association et National Council on Measurement Used in Education (1954). Technical Recommendations for Psychological Tests and Diagnostic Techniques. *Psychological Bulletin*, *51*(2), 1-38.

Andrich, D., Humphry, S. M. et Marais, I. (2012). Quantifying Local, Response Dependence Between Two Polytomous Items Using the Rasch Model. *Applied Psychological Measurement*, *36*(4), 309–324.

Artus, F. et Demeuse, M. (2008). Évaluer les productions orales en français langue étrangère (FLE) en situation de test. Étude de la fidélité inter-juges de l'épreuve d'expression orale du Test d'Évaluation du Français (TEF) de la Chambre de Commerce et d'Industrie de Paris (CCIP). *Les Cahiers des Sciences de l'Éducation*, 25-26, 131-151.

Artus, F., Demeuse, M., Maréchal, M., Casanova, D., Crendal, A., Desroches, F. et Holle, A. (2011). Évaluer la compétence écrite en français des étudiants non-francophones en situation académique. *Actes du 23ème colloque de l'Admée-Europe Évaluation et enseignement supérieur*, 1-10.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99-115.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly*, *2*(1), 1-34.

Bachman, L. F., Lynch, B. K. et Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-257.

Bachman, L. F. et Palmer, A. S. (2010). *Language assessment in practice*. Oxford : Oxford University Press.

Barbier, J.-M. (1983). Pour une histoire et une sociologie des pratiques d'évaluation en formation. *Revue française de pédagogie*, *63*, 47-60.

Bartlett, M. S. (1967). Some Remarks on the Analysis of Time-Series. *Biometrika*, *54*(1-2), 25-38.

Bejar, I. I. (2012). Rater Cognition: Implications for Validity. *Educational Measurement: Issues and Practice*, *31*(3), 2-9.

Béland, S. (2015). *Étude comparative de nouveaux indices de détection de réponses inappropriées*. (Thèse de doctorat non publiée). Université du Québec à Montréal.

Belcher, J., Hampton, J. S. et Tunnicliffe Wilson, G. (1994). Parameterization of Continuous Time Autoregressive Models for Irregularly Sampled Time Series Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*(1), 141-155.

Bernardin, H. J., LaShells, M. B., Smith, P. C. et Alvares, K. M. (1976). Behavioral Expectation Scales: Effects of Developmental Procedures and Formats. *Journal of Applied Psychology*, *61*(1), 75-79.

Bickhard, M. H. (2001). The Tragedy of Operationalism. *Theory & Psychology*, *11*(1), 35-44.

Bonk, W. J. et Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110.

Boswell, J. F., Anderson, L. M. et Barlow, D. H. (2014). An Idiographic Analysis of Change Processes in the Unified Transdiagnostic Treatment of Depression. *Journal of Consulting and Clinical Psychology*, *82*(6), 1060-1071.

Box, G. E. P., Jenkins, G. M. et Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control* (3^{e} éd.). Englewood Cliffs, NJ : Prentice Hall.

Brock, W. A., Dechert, W. D. et Sheinkman J. A. (1987). *A Test of Independence Based on the Correlation Dimension* (SSRI no. 8702). Madison : Department of Economics, University of Wisconsin.

Brockwell, P. J. et Davis, R. A. (2002). *Introduction to Time Series and Forecasting* (2^{e} éd.). New York : Springer.

Brooks, R. L. (2013). *Comparing native and non-native raters of US federal government speaking tests*. (Thèse de doctorat non publiée). Georgetown University, Washington, DC.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Buckingham, B. R., McCAll, W. A., Otis, A. S., Rugg, H. O., Trabue, M. R. et Courtis S. A. (1921). Report of the Standardization Committee. *Journal of Educational Research*, *4*(1), 78-80.

Caban, H. L. (2003). Rater Group Bias In The Speaking Assessment Of Four L1 Japanese ESL Students. *Second Language Studies*, *21*(2), 1-44.

Cai, H. (2012). *Weighting Patterns and Rater Variability in an English as a Foreign Language Speaking Test*. (Thèse de doctorat non publiée). University of California, Los Angeles.

Carey, M. D., Mannell, R. H. et Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219.

Casanova, D. et Demeuse, M. (2011). Analyse des différentes facettes influant sur la fidélité de l'épreuve d'expression écrite d'un test de français langue étrangère. *mesure et évaluation en éducation, 34*(1), 25-53.

Casanova, D. et Demeuse, M. (2016). Évaluateurs évalués : évaluation diagnostique des compétences en évaluation des correcteurs d'une épreuve d'expression écrite à forts enjeux. *mesure et évaluation en éducation*, *39*(3), 59-96.

Chiland, C. (2004). L'examen psychologique. Dans S. Lebovici, R. Diatkine et M. Soulé (dir.) *Nouveau traité de psychiatrie de l'enfant et de l'adolescent* (2e éd.) (p. 563 à 579). Paris : Presses Universitaires de France.

Cole, J. S. et Osterlind, S. J. (2008). Investigating Differences Between Low- and High-Stakes Test Performance on a GeneralEducation Exam. *The Journal of General Education*, *57*(2), 119-130.

Collins, L. M. et Lanza, S. T. (2010). *Latent class and latent transition analysis*. Hoboken, New Jersey : John Wiley & Sons.

Congdon, P. J. et McQueen, J. (2000).The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement*, *37*(2), 163-178.

Conseil de l'Europe (2005). *Cadre européen commun de référence pour les langues*. Paris : Didier.

Crimmins, G., Nash, G., Oprescu, F., Alla, K., Brock, G., Hickson-Jamieson, B. et Noakes, C. (2016). Can a systematic assessment moderation process assure the quality and integrity of assessment practice while supporting the professional development of casual academics?, *Assessment & Evaluation in Higher Education*, *41*(3), 427-441.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5e éd.). New York : Harper & Row.

Cronbach, L. J. et Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Davis, L. (2012). *Rater expretise in a second language speaking assessment: The influence of training and experience*. (Thèse de doctorat non publiée). University of Hawai'i at Manoa.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117-135.

De Boeck, P. et Wilson, M. (2004). *Explanatory item response models – A generalized linear and nonlinear approach*. New York : Springer.

DeCarlo, L. T., Kim, Y. K. et Johnson, M. S. (2011). A Hierarchical Rater Model for Constructed Responses, with a Signal Detection Rater Model. *Journal of Educational Measurement*, *48*(3), 333-356.

DeCoths, T. A. (1977). An analysis of the external validity and applied relevance of three rating formats. *Organizational Behavior and Human Performance*, *19*(2), 247-266.

de Jong, J. et Linacre, J. M. (1993). Rasch Estimation Methods, Statistical Independence and Global Fit. *Rasch Measurement Transactions*, *7*(2), 296-297.

Desai, M. M. (1965). Practical problems in the assessment of personality in the clinical field. *British Journal of Medical Psychology*, *38*(3), 231-240.

Du, Y. et Brown, W. L. (2000). Raters and Single Prompt-To-Prompt Equating Using The Facets Model In A Writing Performance Assessment. Dans M. Wilson et G. Engelhard, Jr. (dir.), *Objective Measurement: Theory Into Practice, volume 5* (p. 97-111). Stamford, CT: Ablex.

Eckes, T. (2005). Examining Rater Effects in TestDaF Writing and Speaking Performance Assessments: A Many-Facet Rasch Analysis. *Language Assessment Quarterly*, *2*(3), 197-221.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155-185.

Eckes, T. (2009). Many-facet Rasch measurement. Dans S. Takala (dir.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (Section H), Strasbourg, France : Council of Europe/Language Policy Division.

Eckes, T. (2011). *Introduction to Many-Facet Rasch Measurement*. Frankfurt am Main : Peter Lang.

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly*, *9*(3), 270-292.

Edgeworth, F. Y. (1888). The Statistics of Examinations. *Journal of the Royal Statistical Society*, *51*(3), 599-635.

Edgeworth, F. Y. (1890). The Element of Chance in Competitive Examinations. *Journal of the Royal Statistical Society*, *53*(4), 644-663.

Elder, C., Barkhuizen, G., Knoch, U. et von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*(1), 37-64.

Elder, C., Knoch, U., Barkhuizen, G. et von Randow, J. (2005). Individual Feedback to Enhance Rater Training: Does It Work? *Language Assessment Quarterly*, *2*(3), 175-196.

Engelhard Jr., G. (1994). Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement*, *31*(2), 93-112.

Engelhard Jr., G. et Myford, C. (2003). *Monitoring Faculty Consultant Performance in the Advanced Placement English Literature and Composition Program with a Many-Faceted Rasch Model.* (College Board Research Report No. 2003-1 ETS RR-03-01). New York : College Entrance Examination Board.

Erickson, G., Åberg-Bengtsson, L. et Gustafsson, J.-E. (2015). Dimensions of test performance in English as a foreign language in different European settings: a two-level confirmatory factor analytical approach. *Educational Research and Evaluation, 21*(3), 188-208.

Eszter, B. (2007). *An Investigation Of Rater And Rating Scale Interaction In The Validation Of The Assessment Of Writing Performance.* (Thèse de doctorat non publiée). Université Eötvös Loránd, Budapest.

Evans, M. K. (2003). *Practical Business Forecasting.* Oxford : Blackwell Publishers.

Fahim, M. et Bajani, H. (2011). The Effects of Rater Training on Raters' Severity and Bias in Second Language Writing Assessment. *Iranian Journal of Language Testing, 1*(1), 1-16.

Feest, U. (2005). Operationism in psychology: what the debate is about, what the debate should be about. *Journal of the History of the Behavioral Sciences, 41*(2), 131–149.

Fischer, G. H. (1983). Some latent trait models for measuring change in qualitative observations. Dans D. J. Weiss (dir.), *New horizons in testing – Latent trait test theory and computerized adaptive testing.* New York : Academic press.

Fuller, R., Homer, M., Pell, G. et Hallam, J. (2016). Managing extremes of assessor judgment within the OSCE. *Medical Teacher.* http://dx.doi.org/10.1080/0142159X.2016.1230189

Girard, Y. (2011). *Séries chronologiques à une et plusieurs variables : synthèse des méthodes classiques et modèles à base de copules.* (Mémoire de maîtrise non publié). Université du Québec à Trois-Rivières.

Guilford, J. P. (1954). *Psychometric methods* (2e éd.). New York : McGraw-Hill.

Hamaker, E. L., Grasman, R. P. P. P. et Kamphuis, J. H. (2016). Modeling BAS Dysregulation in Bipolar Disorder: Illustrating the Potential of Time Series Analysis. *Assessment, 23*(4), 436-446.

Han, Q. (2016). Rater Cognition in L2 Speaking Assessment: a Review of the Literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *16*(1), 1-24.

Harding, L. (2014). Communicative Language Testing: Current Issues and Future Research. *Language Assessment Quarterly*, *11*(2), 186-197.

Hoskens, M. et Wilson, M. (2001). Real-Time Feedback on Rater Drift in Constructed-Response Items: An Example From the Golden State Examination. *Journal of Educational Measurement*, *38*(2), 121-145.

Hox, J. J. et Boeije, H. R. (2005). Data Collection, Primary vs. Secondary. Dans K. Kempf Leonard (dir.), *Encyclopedia of Social Measurement* (p. 593-599). Boston, Londres, Elsevier.

Hsieh, C.-N. (2011). *Rater Effects in ITA Testing: ESL Teachers' versus American Undergraduates' Judgments of Accentedness, Comprehensibility, and Oral Proficiency.* (Thèse de doctorat non publiée). Michigan State University, Ann Arbor.

Huang, B., Alegre, A. et Eisenberg, A. (2016) A Cross-Linguistic Investigation of the Effect of Raters' Accent Familiarity on Speaking Assessment, *Language Assessment Quarterly*, *13*(1), 25-41.

Huang, B. et Jun, S.-A. (2015). Age Matters, And So May Raters: Rater Differences in the Assessment of Foreign Accents. *Studies in Second Language Acquisition*, *37*(4), 623–650.

Hurvich, C. M. et Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297-307.

Hwang, S. et Valls Pereira, P. L. (2006) Small sample properties of GARCH estimates and persistence. *The European Journal of Finance*, *12*(6-7), 473-494.

Hyndman, R. J. (2016). *forecast* (version 7.2) [Bibliothèque R]. https://github.com/robjhyndman/forecast

Hyndman, R. J. et Athanasopoulos, G. (2014). *Forecasting: principles and practice*. OTexts. Récupéré de https://www.otexts.org/book/fpp

Hyndman, R. J. et Kostenko, A. V. (2007). Minimum sample size requirements for seasonal forecast models. *Foresight*, 6, 12-15.

Johnson, J. S. et Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505.

Kachchaf, R. et Solano-Flores, G. (2012). Rater Language Background as a Source of Measurement Error in the Testing of English Language Learners. *Applied Measurement in Education, 25*(2), 162-177.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1-73.

Kang, O. (2008). *Ratings of L2 oral performance in English: relative impact of rater characteristics and acoustic measures of accentedness*. (Thèse de doctorat non publiée). University of Georgia, Athens.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*, *1*(2), 152-176.

Kassim, N. L. A. (2007, juin). *Exploring Rater Judging Behaviour Using the Many-Facet Rasch Model.* Communication présentée à The Second Biennial International Conference on Teaching and Learning of English in Asia: Exploring New Frontiers (TELiA2), Holiday Villa Beach & Spa Resort, Langkawi, Malaysie.

Kim, H. J. (2011). *Investigating raters' development of rating ability on a second language speaking assessment.* (Thèse de doctorat non publiée). Teachers College, Columbia University, New York.

Kim, Y.-H. (2009). A G-Theory Analysis of Rater Effect in ESL Speaking Assessment. *Applied Linguistics*, *30*(3), 435-440.

Klenowski, V. et Wyatt-Smith, C. (2014). Assessment for Education: Standards, Judgement and Moderation. Thousand Oaks, California : SAGE.

Kline, T. J. B. et Sulsky, L. M. (2009). Measurement and Assessment Issues in Performance Appraisal. *Canadian Psychology*, *50*(3), 161-171.

Knoch, U., Fairbairn, J. et Huisman, A. (2015). *An Evaluation of the Effectiveness of Training Aptis Raters Online* (Rapport de recherche VS/2015/001). British Council, Récupéré le 13 novembre 2016 de : britishcouncil.org/sites/default/files/knoch_fairbairn_and_huisman_0.pdf

Kondo, Y. (2010). Examination of Rater Training Effect and Rater Eligibility in L2 Performance Assessment. *Journal of Pan-Pacific Association of Applied Linguistics*, *14*(2), 1-23.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing*, *19*(1), 3-31.

Korenovska, L. (2013). *An exploration of test taker, rater, and item facets of the writing section of TOEFL using Many-Facet Rasch Measurement*. (Thèse de doctorat non publiée). Fordham University, New York.

Kreiner, S. et Christensen, K. B. (2016). *Exact evaluation of bias in Rasch model residuals*. (Rapport de recherche 07/2). Copenhague : Département de biostatistique, Université de Copenhague.

Kroonenberg, P. M. (2008). *Applied multiway data analysis*. New York, NY: Wiley.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. et Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*(1–3), 159–178.

Landy, F. J. et Farr, J. L. (1980). Performance rating. *Psychological Bulletin*, *87*(1), 72-107.

Laugier, H. et Weinberg, D. (1927). Le Facteur subjectif dans les notes d'examen. *L'année psychologique*, *28*, 236-244.

Laveault, D. et Yerly, G. (2017). Modération statistique et modération sociale des résultats scolaires : approches opposées ou complémentaires? *Mesure et évaluation en éducation*, *40*(2), 91-123.

Leckie, G. et Baird, J.-A. (2011). Rater Effects on Essay Scoring: A Multilevel Analysis of Severity Drift, Central Tendency, and Rater Experience. *Journal of Educational Measurement*, *48*(4), 399-418.

Libman, Z. (2009). Teacher Licensing Examinations – True Progress or an Illusion? *Studies in Educational Evaluation, 35*(1), 7-15.

Lim, G. (2009). *Prompt and rater effects in second language writing and performance assessment*. (Thèse de doctorat non publiée). Michigan State University, Ann Arbor.

Lim, G. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language testing*, *28*(4), 543-560.

Linacre, J. M. (1994). *Many-Facet Rasch Measurement* (2e éd.). Chicago, Illinois : Mesa.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis (PCA)? *Rasch Measurement Transactions*, *12*(2), 636. Récupéré le 17 octobre 2016 de rasch.org/rmt/rmt122m.htm

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardised mean? *Rasch Measurement Transaction*, *16*, 878.

Linacre, J. M. (2009). Local Independence and Residual Covariance: A Study of Olympic Figure Skating Ratings. *Journal of Applied Measurement*, *10*(2), 157-169.

Linacre, J. M. (2014). *Facets computer program for many-facet Rasch measurement*, version 3.71.4. Beaverton, Oregon : Winsteps.com.

Linacre, J. M. (2015). *Winsteps® Rasch measurement computer program*, version 3.90.2. Beaverton, Oregon : Winsteps.com.

Linacre, J. M. (2017a). *Facets computer program for many-facet Rasch measurement User's Guide*, Beaverton, Oregon : Winsteps.com.

Linacre, J. M. (2017b). *Winsteps® Rasch measurement computer program User's Guide*. Beaverton, Oregon : Winsteps.com.

Linacre, J. M. et Wright, B. D. (1989). The "Length" of a Logit. *Rasch Measurement Transactions*, *3*(2), 54-55.

Ljung, G. M. et Box, G. E. P. (1978). On a Measure of a Lack of Fit in Time Series Models. *Biometrika*, *65*(2), 297-303.

Lopes Toffoli, S. F., de Andrade, D. F. et Bornia, A. C. (2016). Evaluation of open items using the many-facet Rasch model, *Journal of Applied Statistics*, *43*(2), 299-316.

Lumley, T. et McNamara, T. F. (1995). Rater characteristics and rater bias: implications for training. *Language Testing*, *12*(1), 54-71.

MacCorquodale, K.et Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, *55*, 95-107.

Mallinson, T., Pape, T. et Guernon, A. (2016, mai). *Accounting For Rater Severity/Leniency In Endpoint Measures Of Recovery Of Consciousness In Adults With Severe Traumatic Brain Injury*. Communication présentée au ISPOR 21st Annual International Meeting, Washington, D. C.

Martin, J. (2002). Aux origines de la « science des examens » (1920-1940). *Histoire de l'éducation*, *94*, 177-199.

Maul, A. et McGrane, J. (2017). As Pragmatic as Theft Over Honest Toil: Disentangling Pragmatism From Operationalism. *Measurement: Interdisciplinary Research and Perspectives*, *15*(1), 2-4.

McClellan, C. A. (2010). Constructed-Response Scoring: Doing It Right. *R & D Connections*, (13), 1-7.

McKinley, D. et Boulet, J. (2004). Detecting Score Drift in a High-Stakes Performance-Based Assessment. *Advances in Health Sciences Education, 9*(1), 29-38.

McManus, I. C., Thompson, M. et Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education*, 6-42.

Meier, V. (2014). Evaluating Rater and Rubric Performance on a Writing Placement Exam. *Second Language Studies*, *31*(1), 47-101.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, *50*(9), 741-749.

Ministère de l'Enseignement supérieur, de la Recherche, de la Science et de la Technologie. (2013). Épreuve uniforme de français, langue d'enseignement et littérature : Toute l'information de A à Z. Québec : Gouvernement du Québec.

Ministère de l'Immigration, Diversité et Inclusion. (2018). Connaissances en français et en anglais pour les candidats du Programme régulier des travailleurs qualifiés. Dans *Connaissances linguistiques*. Récupéré de https://www.immigration-quebec.gouv.qc.ca/fr/immigrer-installer/travailleurs-permanents/conditions-requises/connaissances-linguistiques.html

Myford, C. M. (2012). Rater Cognition Research: Some Possible Directions for the Future. *Educational Measurement: Issues and Practice*, *31*(3), 48-49.

Myford, C. M., Marr, D. B. et Linacre, J. M. (1996). *Reader Calibration and its Potential Role in Equating for the Test of Written English*. (Rapport de recherche RR-95-40). Princeton, NJ, États-Unis : Educational Testing Service.

Myford, C. M. et Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement*, *4*(4), 386-422.

Myford, C. M. et Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Nason, G. (2016). *locits* (version 1.7.1) [Bibliothèque R]. https://people.maths.bris.ac.uk/~magpn/

Newton, P. E. et Shaw, S. D. (2014). *Validity in Educational & Psychological Assessment*. Thousand Oaks, California : Sage.

Norman, W. T. et Goldberg, L. R. (1966). Raters, ratees, and randomness in personality structure. *Journal of Personality and Social Psychology*, *4*(6), 681-691.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing*, *19*(2), 169-192.

Pankratz, A. (1983). *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*. New York : John Wiley & Sons.

Park, Y. S. (2011). *Rater drift in constructed response scoring via latent class signal detection theory and item response theory*. (Thèse de doctorat non publiée). Columbia University.

Prieto, G. et Nieto, E. (2014). Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement. *Psicológica*, *35*, 385-397.

R Core Team (2013). *R: A language and environment for statistical computing* (version 3.3.1). R Foundation for Statistical Computing, Vienne, Autriche.

Raîche, G., Magis, D., Blais, J.-G. et Brochu, P. (2013). Taking atypical response patterns into account. Dans M. Simon, K. Ercikan et M. Rousseau (dir.), *Improving large-scale assessments in education – Theory, issues and practice*. Boca Raton, California: Routledge.

Ramsay, J. O. et Silverman, B. W. (2005). *Functional Data Analysis* (2e éd.). New York : Springer.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogische Institut.

Rehfeld, K., Marwan, N., Heitzig, J. et Kurths, J. (2011). Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics*, *18*, 389–404.

Ryan, K. (2002). Assessment Validation in the Context of High-Stakes Assessment. *Educational Measurement: Issues and Practice*, *21*(1), 7-15.

Saal, F. E., Downey, R. G. et Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

Said, S. E. et Dickey, D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika*, *71*(3), 599–607.

Savchev, D. et Nason, G. (2015). *hwwntest* (version 1.3) [Bibliothèque R]. https://cran.r-project.org/web/packages/hwwntest/index.html

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493.

Schönbrodt, F. D. et Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609-612.

Schriesheim, C. A., Kinicki, A. J. et Schriesheim, J. F. (1979). The effect of leniency on leader behavior descriptions. *Organizational Behavior and Human Performance*, *23*(1), 1-29.

Seol, H. (2016). Using the Bootstrap Method to Evaluate the Critical Range of Misfit for Polytomous Rasch Fit Statistics. *Psychological Reports*, *118*(3), 937-956.

Shaw, S. (2002). The effect of training and standardisation on rater judgement and inter-rater reliability. *Research Notes*, *8*, 13–17. Récupéré de : www.cambridgeesol.org/rs_notes/rs_nts8.pdf

Shin, Y. (2017). *Time Series Analysis in the Social Sciences*. Oakland, CA : University of California Press.

Smith, R. M. (1988). The distributional properties of Rasch residuals. *Education and Psychological Measurement*, *48*(3), 657-667.

Smith, R. M. (1991). The distributional properties of Rasch item fit statistics. *Education and Psychological Measurement*, *51*(3), 541-565.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G. et Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *8*(33). DOI : 10.1186/1471-2288-8-33

Snippe, E., Bos, E. H., van der Ploeg, K. M., Sanderman, R., Fleer, J. et Schroevers, M. J. (2015). Time-Series Analysis of Daily Changes in Mindfulness, Repetitive Thinking, and Depressive Symptoms During Mindfulness-Based Treatment. *Mindfulness*, 6, 1053-1062.

Stobart, G. et Eggen, T. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education : Principles, Policy & Practice*, *19*(1), 1-6.

Stray, C. (2001). The Shift from Oral to Written Examination: Cambridge and Oxford 1700–1900. *Assessment in Education: Principles, Policy & Practice*, *8*(1), 33-50.

Sturman, M. C. (2003). Searching for the Inverted U-Shaped Relationship Between Time and Performance: Meta-Analyses of the Experience/Performance, Tenure/Performance, and Age/Performance Relationships. *Journal of Management*, *29*(5), 609–640.

Taylor, E. K. et Hastman, R. (1956). Relation of Format and Administration to Characteristics of Graphic Rating Scales. *Personnel Psychology*, *9*(2), 181-206.

Teräsvirta, T., Lin, C.-F. et Granger, C. W. J. (1993). Power of the Neural Network Linearity Test. *Journal of Time Series Analysis*, *14*(2), 209-220.

Thibodeau, K. (2011). *Application de la méthodologie Box-Jenkins aux séries du ministère de la Santé*. (Mémoire de maîtrise non publié). Université du Québec à Trois-Rivières.

Trapletti, A., Hornik, K. et LeBaron, B. (2016). *tseries* (version 0.10-35) [Bibliothèque R]. https://cran.r-project.org/web/packages/tseries/index.html

Upshur, J. A. et Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing*, *16*(1), 82-111.

Várkonyi, Z. (2012, juillet). Un opérateur à peine franc-comtois. Actes du colloque *CMLF année – 3ème Congrès Mondial de Linguistique Française*, Lyon, du 4 au 7 juillet (p. 2011-2026). EDP Sciences (www.linguistiquefrancaise.org). DOI 10.1051/shsconf/20120100200

Wang, Z. et Yao, L. (2013). *The Effects of Rater Severity and Rater Distribution on Examinees' Ability Estimation for Constructed-Response Items*. (Rapport de recherche ETS RR-13-23). Princeton, New Jersey, États-Unis : Educational Testing Service.

Wang, W.-C., Su, C.-M. et Qiu, X.-L. (2014). Item Response Models for Local Dependence Among Multiple Ratings. *Journal of Educational Measurement*, *51*(3), 260-280.

Wei, J. et Llosa, L. (2015). Investigating Differences Between American and Indian Raters in Assessing TOEFL iBT Speaking Tasks. *Language Assessment Quarterly*, *12*(3), 283-304.

Weigle, S. C. (1994). Effects of training on raters of ESL composition. *Language Testing*, *11*(2), 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263-287.

Wesolowski, B. C. (2016). Exploring rater cognition: A typology of raters in the context of music performance assessment. *Psychology of Music* DOI : 10.1177/0305735616665004

Wickham, P. (2015). *normwhn.test* (version 1.0) [Bibliothèque R]. https://cran.r-project.org/web/packages/normwhn.test/index.html

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, *10*(3), 305-319.

Wilson, M. et Case, H. (1997). *An Examination of Variation in Rater Severity Over Time: A Study in Rater Drift.* (Rapport de recherche). University of California, Berkeley.

Wilson M. et Hoskens M. (2001). The Rater Bundle Model, *Journal of Educational and Behavioral Statistics*, *26*(3), 283-306.

Wind, S. A. et Engelhard Jr., G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*(4), 278-299.

Winke, P., Gass, S. et Myford, C. (2011). *The Relationship Between Raters' Prior Language Study and the Evaluation of Foreign Language Speech Samples.* (Rapport de recherche TOEFL iBT–16). Princeton, New Jersey, États-Unis : Educational Testing Service.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*(3), 150-173.

Wolfe, E. W. et McVay, A. (2010). *Rater Effects as a Function of Rater Training Context.* (Rapport de recherche). Pearson Assessment.

Wolfe, E. W. et McVay, A. (2012). Application of Latent Trait Models to Identifying Substantively Interesting Raters. *Educational Measurement: Issues and Practice*, *31*(3), 31–37.

Wolfe, E. W., Myford, C. M., Engelhard Jr., G. et Manalo, J. R. (2007). *Monitoring Reader Performance and DRIFT in the AP® English Literature and Composition Examination Using Benchmark Essays.* (Rapport de recherche no 2007-2). College Board, New York : College Board.

Wolfe, E. W. et Song, T. (2015). Methods for Monitoring and Document Rating Quality. Dans H. Jiao et R. W. Lissitz (dir.), *The Next Generation of Testing: Common Core Standards, Smarter-Balanced, PARCC, and the Nationwide Testing Movement* (p. 107-143), Baltimore : Information Age Publishing.

Wright, B. D. (1998). Estimating Rasch measures for extreme scores. *Rasch Measurement Transactions*, *12*(2), 632-633.

Wuertz, D. et Chalabi, Y. (2015). *fNonlinear* (version 3010.78) [Bibliothèque R]. https://cran.r-project.org/web/packages/fNonlinear/index.html

Wu, M. et Adams, R. J. (2013). Properties of Rasch residual fit statistics. *Journal of Applied Measurement*, *14*(4), 339-355.

Wu, S. M. et Tan, S. (2016). Managing rater effects through the use of FACETS analysis: the case of a university placement test. *Higher Education Research & Development*, *35*(2), 380-394.

Xi, X. (2010). How do we go about investigating test fairness. *Language testing*, *27*(2), 147-170.

Xi, X. et Mollaun, P. (2009). *How Do Raters From India Perform in Scoring the TOEFL iBT™ Speaking Section and What Kind of Training Helps?* (Rapport de recherche TOEFLiBT-11). Princeton, New Jersey, États-Unis : Educational Testing Service.

Yen, S. H., Ochieng, C., Michaels, H. et Friedman, G. (2005, avril). *The Effect of Year-to-Year Rater Variation on IRT Linking*. Communication présentée à la rencontre annuelle de l'American Educational Research Association, Montréal.

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing*, *27*, 37-53.

Zhang, Y. et Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31-50.

Zumbach, G. (2000). The Pitfalls in Fitting Garch (1,1) Processes. Dans C. L. Danis (dir.), *Advances in Quantitative Asset Management* (p. 179-200), Norwell : Kluwers Academic Publishers.

GLOSSAIRE

Corrélation désatténuée : corrélation linéaire de Pearson entre deux variables, *X* et *Y*, qui corrige la valeur obtenue du coefficient de corrélation en tenant compte de l'erreur de mesure de *X* et *Y*. Le coefficient de corrélation est divisé par la racine carrée du produit des indices de fidélité de *X* et *Y*. Traduction de « *disattenuated correlations* »

Dérive temporelle de la sévérité : traduction libre des expressions anglaises « *rater drift* » ou « *severity drift* ». Désigne les changements de niveau de sévérité d'un examinateur au fil du temps, l'ampleur de ces changements pouvant varier d'un temps à l'autre. A généralement, dans la littérature en anglais, une connotation négative, ce qui explique le choix du mot « dérive » en français, lui aussi négativement connoté.

Devis en spirale incomplet lié : traduction libre de l'expression anglaise « *connected incomplete spiral design* », que l'on trouve dans la littérature portant sur le modèle de Rasch. Expression équivalente à « échantillonnage matriciel ». Désigne un devis dans lequel tous les éléments d'un ensemble sont liés, puisque tous les éléments appartiennent à des sous-ensembles qui se chevauchent (ils partagent au moins un élément commun) et incluent tous les éléments de l'ensemble.

Échelle d'appréciation : Traduction de l'expression anglaise « *rating scale* ». Il y a, en français, deux expressions équivalentes, soit « échelle de notation » et « échelle d'appréciation ». La seconde a été retenue afin d'insister sur l'importance du jugement évaluatif (l'appréciation d'une performance), par rapport au « simple » acte d'accorder une note.

Effets de l'examinateur : expression très rare, mais néanmoins attestée en français, en psychiatrie ou en sociolinguistique (Chiland, 2004 ; Várkonyi, 2012). Équivalent français de l'anglais « *rater effects* », qui désigne l'ensemble des comportements de notation pouvant affecter la note accordée par un examinateur, sans que ces comportements ne soient en lien direct avec la qualité de l'objet évalué. Sa connotation neutre est préférée à des expressions péjoratives comme « biais de l'examinateur » ou « erreurs de l'examinateur », puisque les recherches ont montré que ces effets étaient très répandus.

Examen à forts enjeux : traduction de l'expression anglaise « *high stakes test* ». Deux expressions équivalentes existent en français, soit « évaluation à forts enjeux » ou « examen à forts enjeux ». La seconde a été retenue parce qu'elle met de l'avant le

caractère organisé et circonscrit de l'examen, par rapport à son hyperonyme « évaluation ». L'expression « à forts enjeux » désigne une situation d'évaluation ayant des conséquences *immédiates* considérées importantes par le candidat. C'est-à-dire que l'examen en lui-même, considéré en isolation, mène à une décision affectant le candidat. Par exemple : l'accès à un ordre professionnel ou à un programme/institution d'enseignement, une promotion au travail, la diplomation ou la certification, la reconnaissance des compétences langagières pour l'immigration... Cela différencie ces situations des situations d'évaluation à « faibles enjeux », où l'évaluation est importante, mais ne mène pas par elle-même à une décision immédiate affectant significativement le candidat. C'est la définition habituelle qui se trouve dans la littérature anglophone (Cole et Osterlind, 2008).

Examinateur : traduction retenue de « *rater* » et de « *judge* », qui sont les deux termes utilisés dans la littérature anglophone. Ce terme a été choisi pour insister sur l'aspect formel des situations d'évaluation concernées par cette thèse et pour montrer que l'examinateur est un maillon subordonné à toute une chaîne évaluative et non un évaluateur autonome et indépendant. Est préféré aux termes « juge » ou « évaluateur », que l'on trouve couramment en français. Le premier est trop fortement connoté par son caractère juridique ou sportif et le second est considéré plus vaste que le terme « examinateur », c'est un hyperonyme d'« examinateur ». Un enseignant est, par exemple, un évaluateur lorsqu'il évalue ses élèves, mais il n'est pas un « examinateur » au sens retenu par cette thèse.

Facette factice : traduction libre de l'expression anglaise « *dummy facet* », utilisée dans la littérature sur le modèle de Rasch à multifacettes. Désigne une facette dont les éléments ne servent pas à estimer les mesures des effets principaux (niveau d'habileté, difficulté des items, sévérité des examinateurs...), mais bien à estimer des interactions entre une variable jugée importante par l'analyste (sociodémographique, date, langue parlée, expérience professionnelle...) et l'un des effets principaux.